Scaling Reliability at Dropbox

Our journey towards a distributed ownership model



Sat Kriya Khalsa

Little bit about me…

- Operations manager & SRE manager at previous gig
- Technical Program Manager at Dropbox for ~1 year
- Member of the Reliability Team

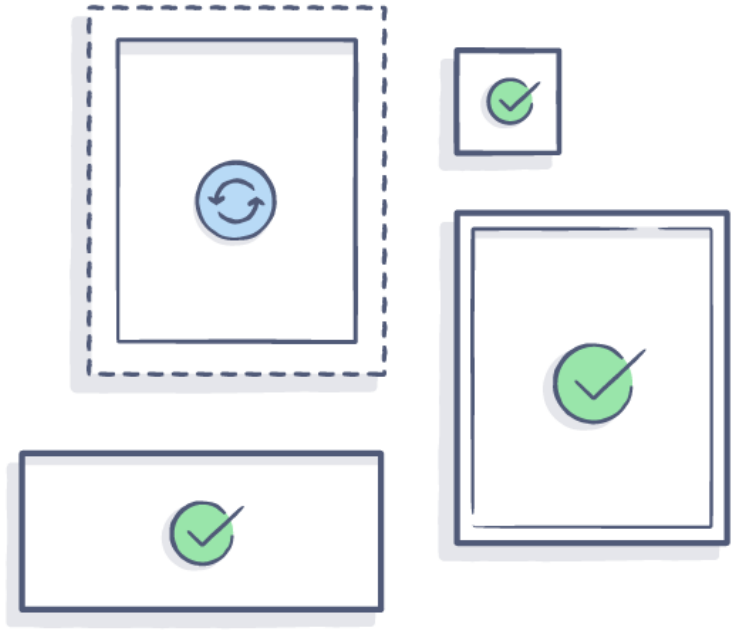


What am I going to cover?



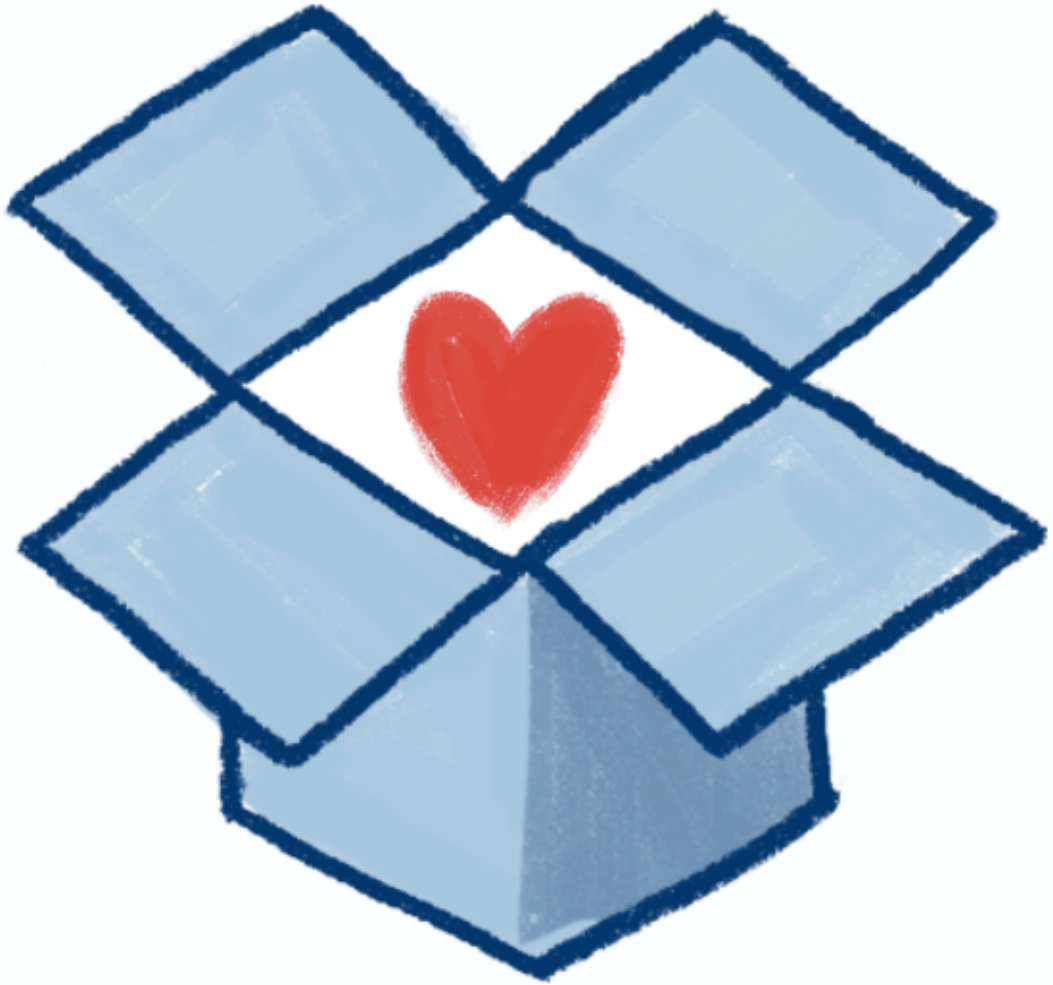
- Background
- Changes we've made
- Where we're going
- How you can get started

Why does reliability matter to Dropbox?





#beworthyoftrust

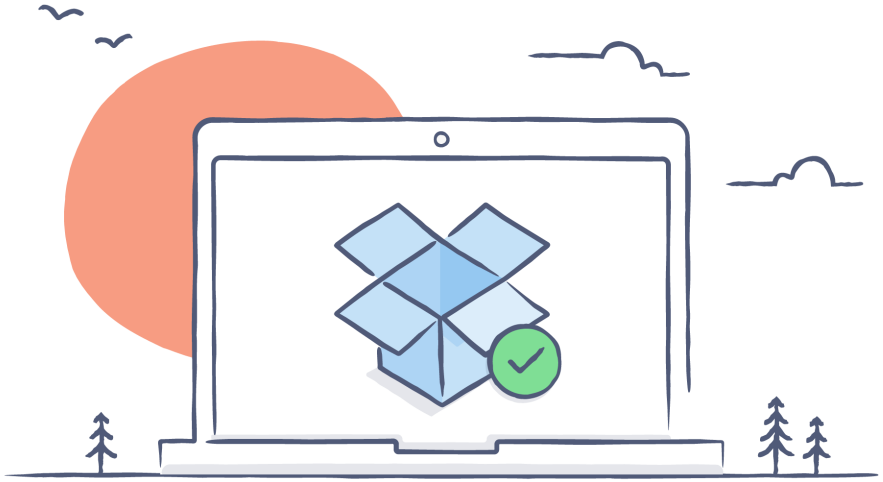


Hold on, what is reliability?



Durability

Data must be not be lost



Availability

Data must be served correctly



Performance

Data must arrive in a timely manner



Enough of the intro stuff!



In the beginning…





Hyper-growth!



What does that mean for infra?



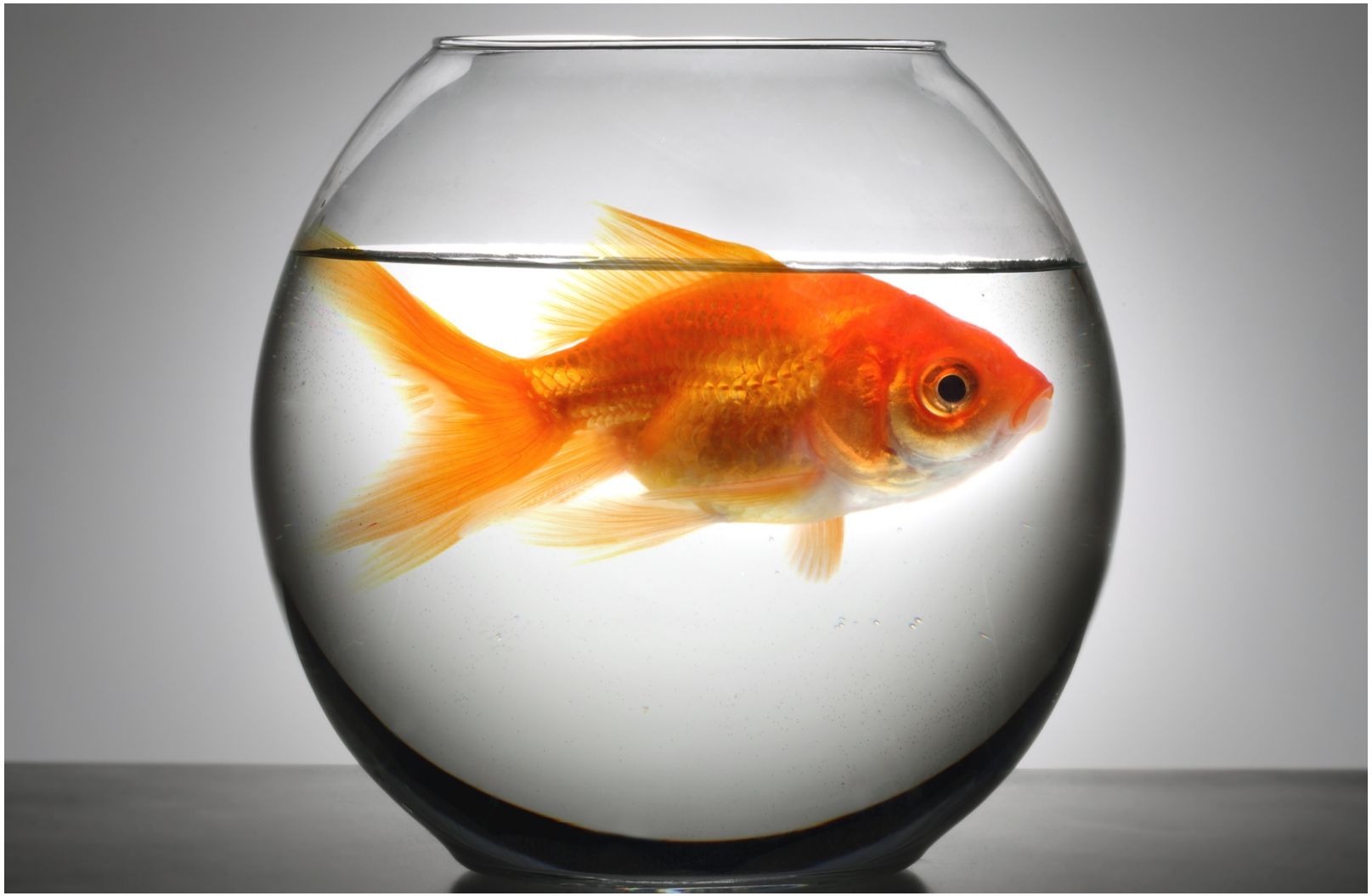
Trust the A-team to get it done!



Collaboration was key



What changed?





Inflection point

- Size
- Complexity



Can’t just throw people at it



Open questions

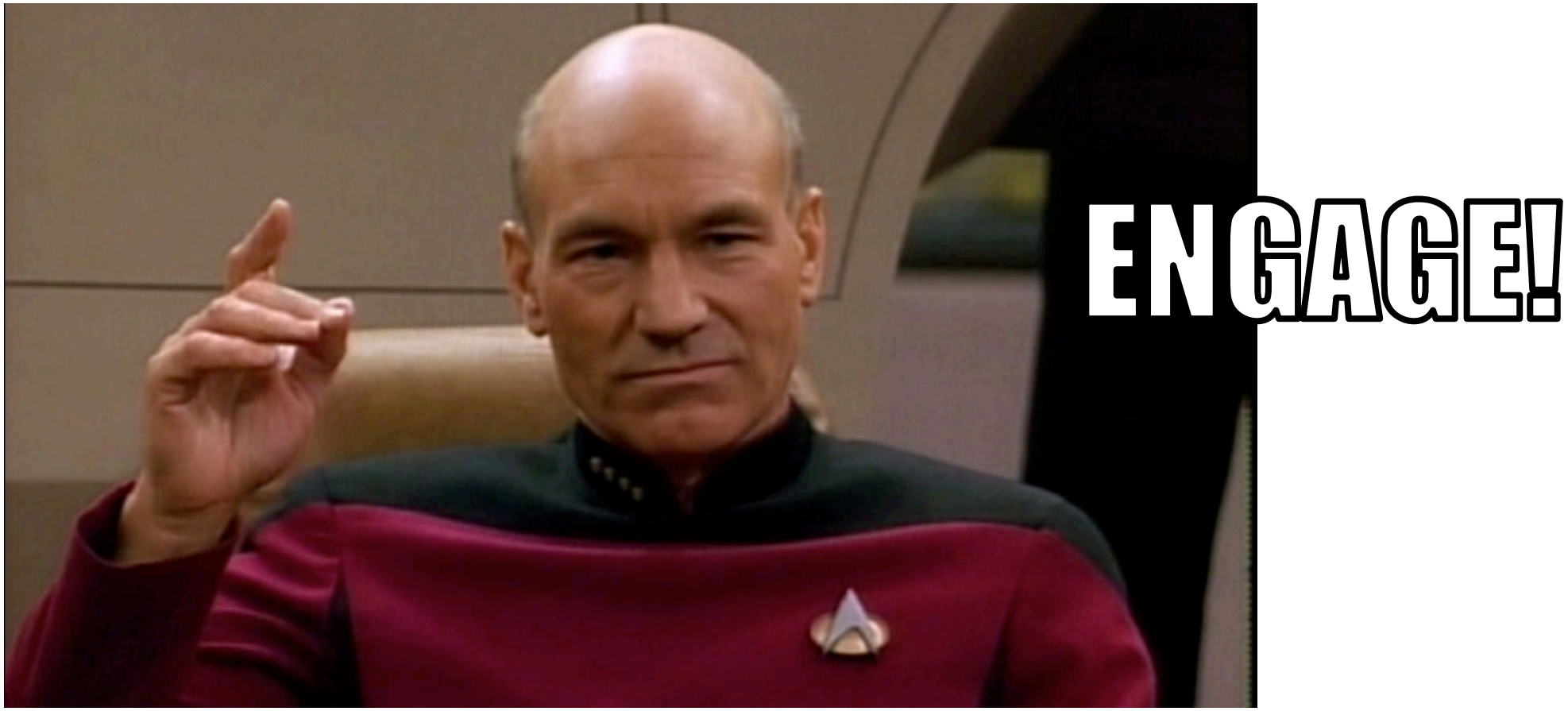
- Which services are impacting our users?
- Why are they breaking?
- Have we fixed it?
- Can we be faster next time?



What did we decide?



SEV Process



What is a SEV?







Types of SEVs



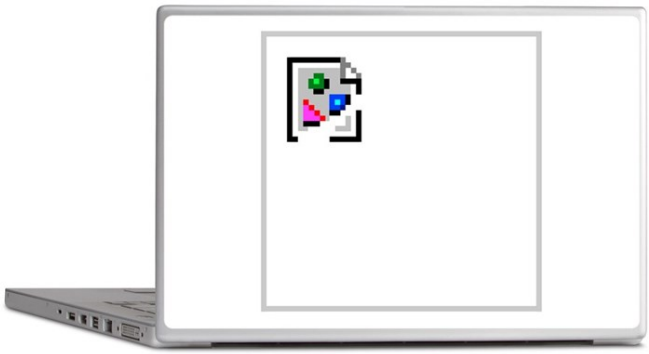
Data Loss



Security Risk



Availability Drop



Product Misbehavior

What is the SEV process?



Incident Management



Postmortem



Reliability Review



Action Items

What is the SEV process?



Incident Management



Postmortem



Reliability Review



Action Items

Preparation is critical

Plans are worthless, but planning is everything.

Eisenhower, addressing the National Defense Executive Reserve



What is the nature of your emergency?



SEV3 - *Minimal service impact*



SEV2 - *Moderate service impact*



SEV1 - *Significant service impact*



SEV0 - *Critical service impact*



Invest in trained responders



What is the SEV process?



Incident Management



Postmortem



Reliability Review



Action Items

Impact statistics

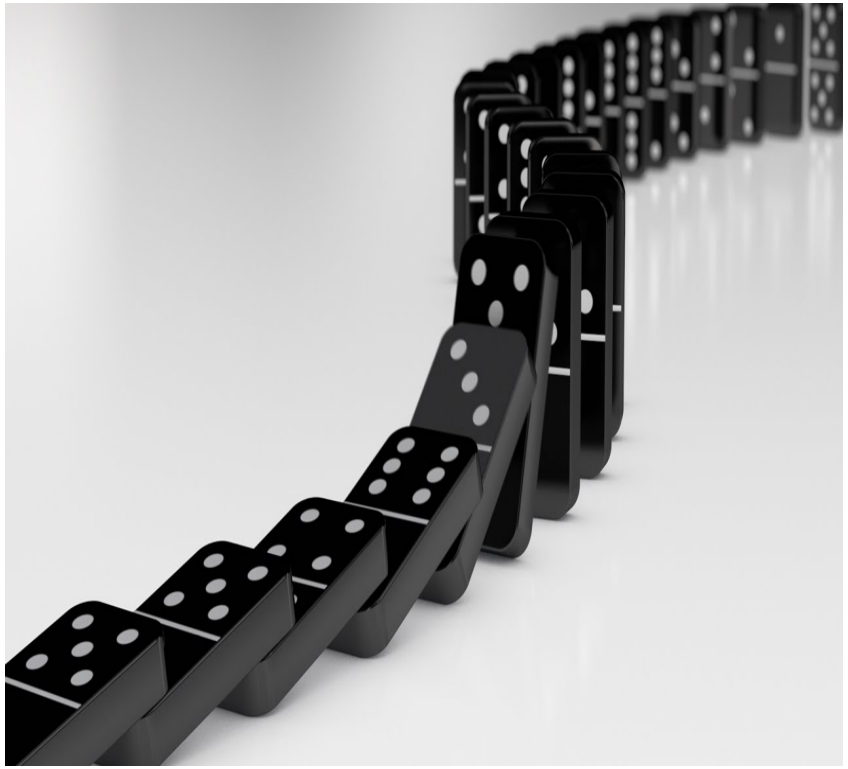


Users Impacted

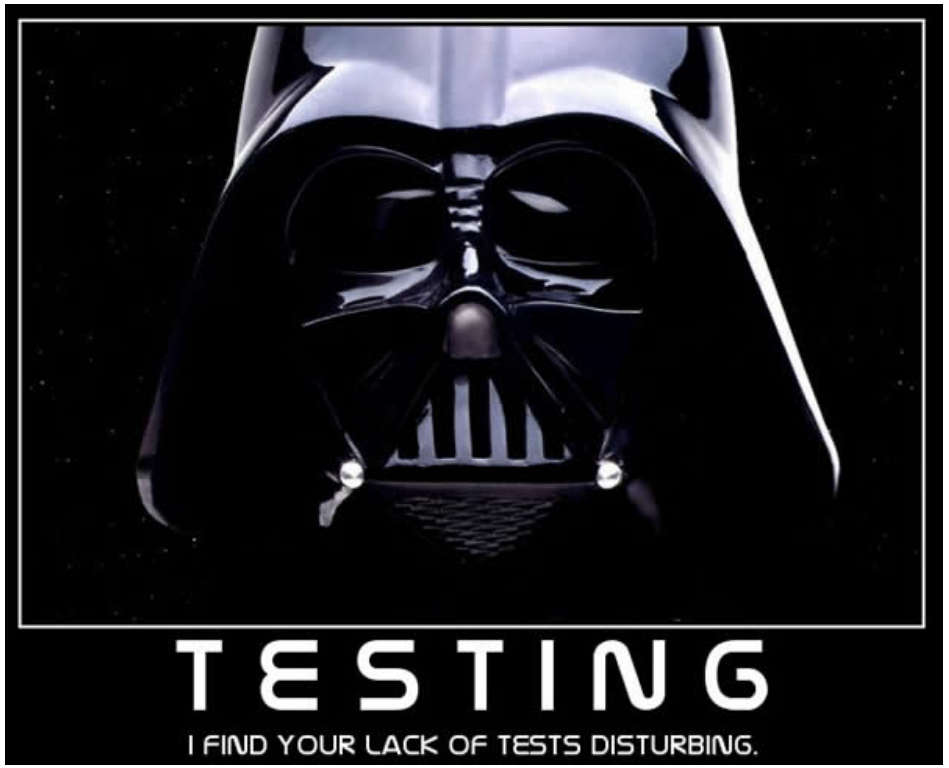


Time to Recover

Failure modes



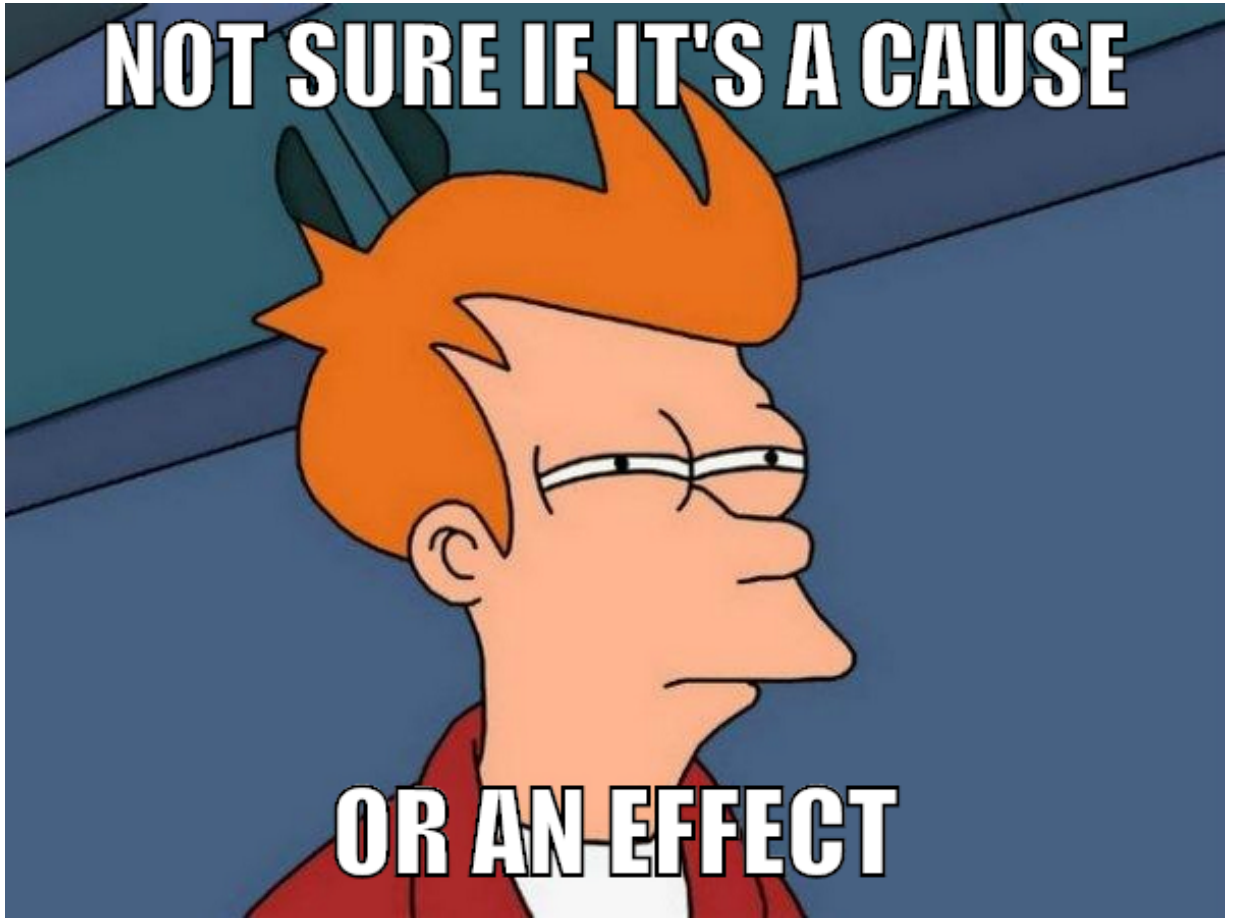
Systems Issues



Process Issues



Dependencies





What is the SEV process?



Incident Management



Postmortem



Reliability Review



Action Items

Meeting structure

Role	Responsibilities
People Involved + IMOC & TLOC	Present factual/technical details of the SEV along with the AIs
Senior Engineer(s)	Validate technical direction, look for gaps in strategy, ensure changes make sense within the global context
Director(s)	**Oversight**. Ensure the prioritization of AIs. Ensure AIs are complete and accurate
Process Owner(s)	Accountability and meeting owner



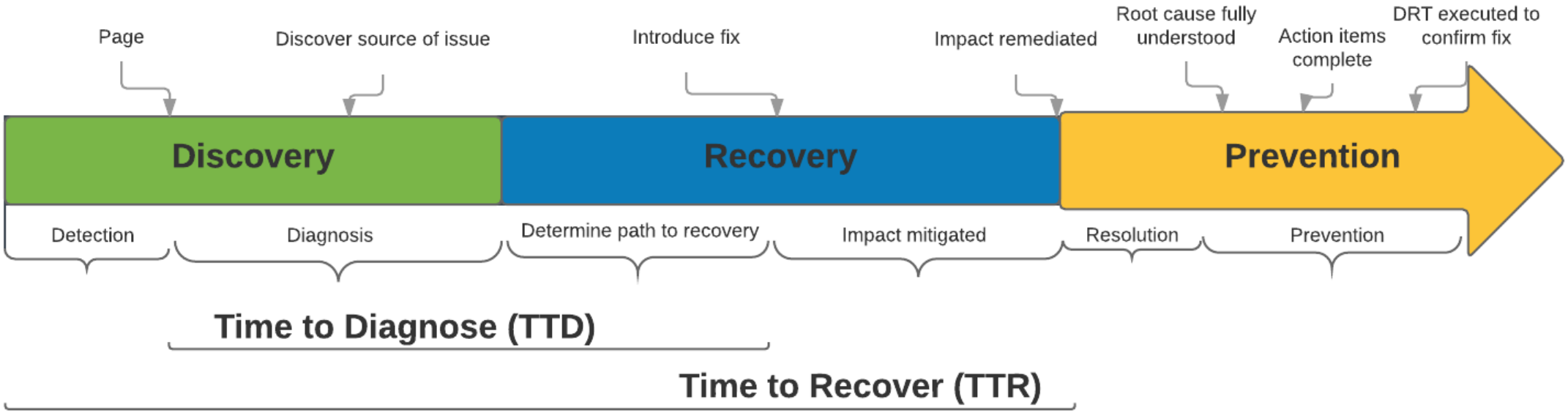
First rule of reliability reviews…

BE

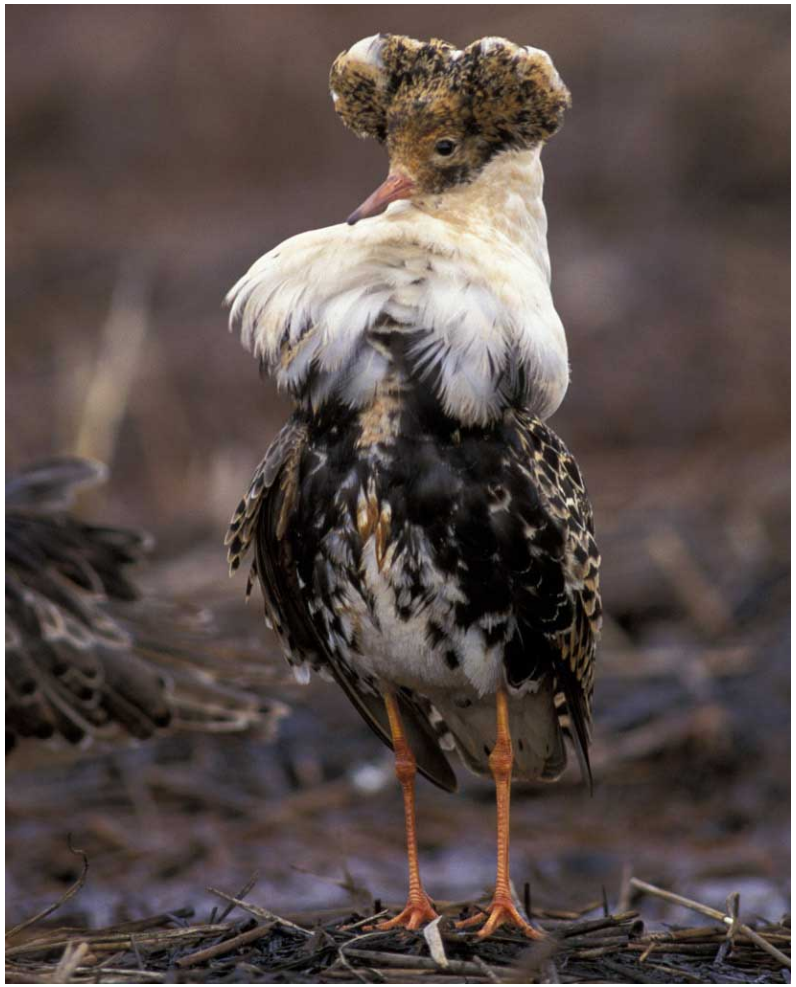
BLAMELESS



Meeting format



Cruel Ruff – Critical feature broken on iOS





Discovery [26 days]

Detection - 25 days

- Issue went undetected by both internal and beta users. First discovered externally. Customer Support was notified and contacted engineering. Investigation revealed that all customers running the new build were affected and filed a SEV.

Diagnosis - 2.5 hours

- Several engineers across different teams investigated and were able to and identify the root cause.

Action Items for Faster Discovery

- Add QA tests on Android/iOS (manual testing)
- Add at least one UI test on Android/iOS (automated testing)
- Setup alert for tracking significant app changes for Android & iOS
- Teach/empower customer support to file SEVs directly



Strange new world!



Tammy Butow

SRE Manager for Databases & Storage

"SEV0 reviews with [VP] it's my favorite meeting –

I always learn so much"

What is the SEV process?



Incident Management



Postmortem



Reliability Review



Action Items

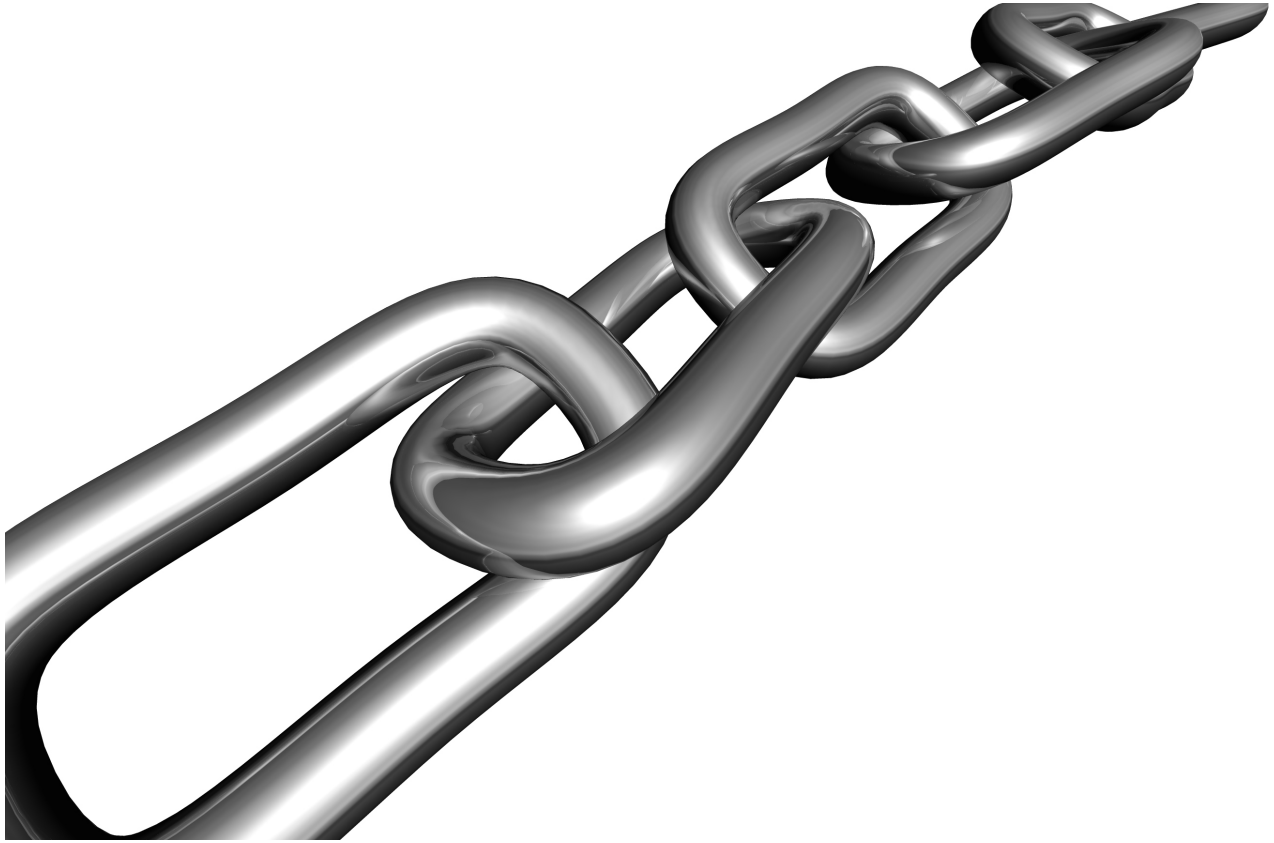
Schrödinger's tasks



Output from the process?



Which services are impacting our users?





Critical System

Any system that can impact availability or durability



Why is it breaking?





Expect failure!

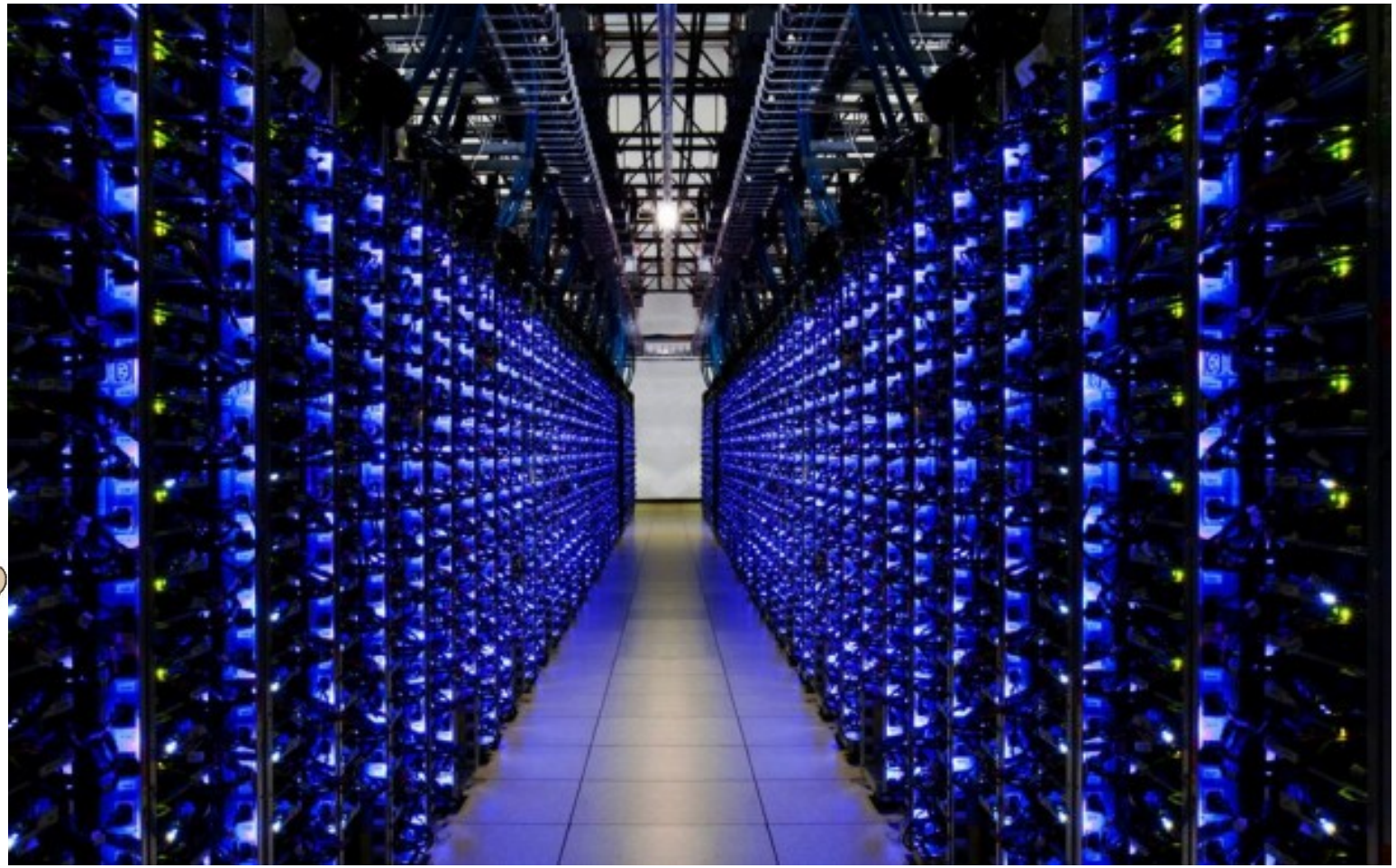
Example cluster with 10's of thousands of machines you will see the following in the **first year:**

- ~0.5 overheating (power down most machines in <5 mins, ~1-2 days to recover)
- ~1 PDU failure (~500-1000 machines suddenly disappear, ~6 hours to come back)
- ~1 rack-move (plenty of warning, ~500-1000 machines powered down, ~6 hours)
- ~1 network rewiring (rolling ~5% of machines down over 2-day span)
- ~20 rack failures (40-80 machines instantly disappear, 1-6 hours to get back)
- ~5 racks go wonky (40-80 machines see 50% packetloss)
- ~8 network maintenances (4 might cause ~30-minute random connectivity losses)
- ~12 router reloads (takes out DNS and external vips for a couple minutes)
- ~3 router failures (have to immediately pull traffic for an hour)
- ~dozens of minor 30-second blips for dns
- ~1000 individual machine failures
- ~thousands of hard drive failures slow disks, bad memory, misconfigured machines, flaky machines, etc...



Bringing chaos back!



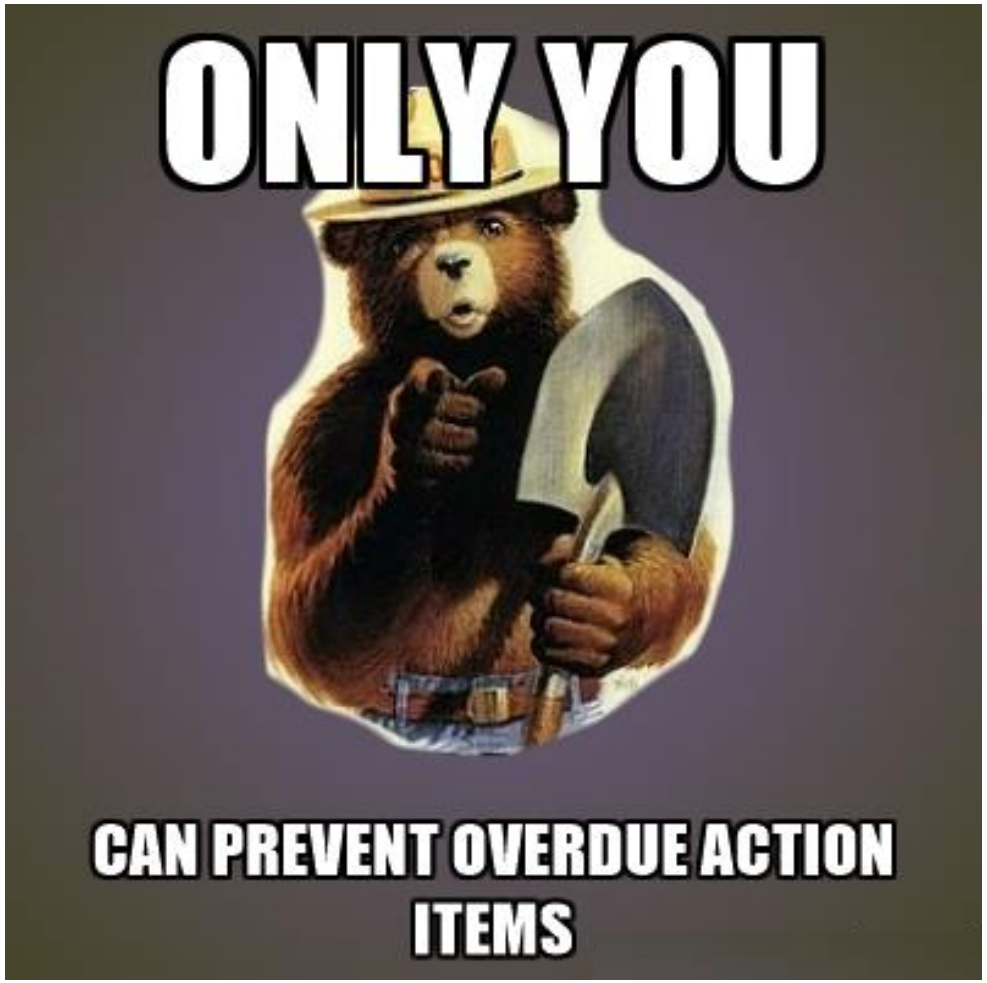


Test early and often! Moar DRT(s)



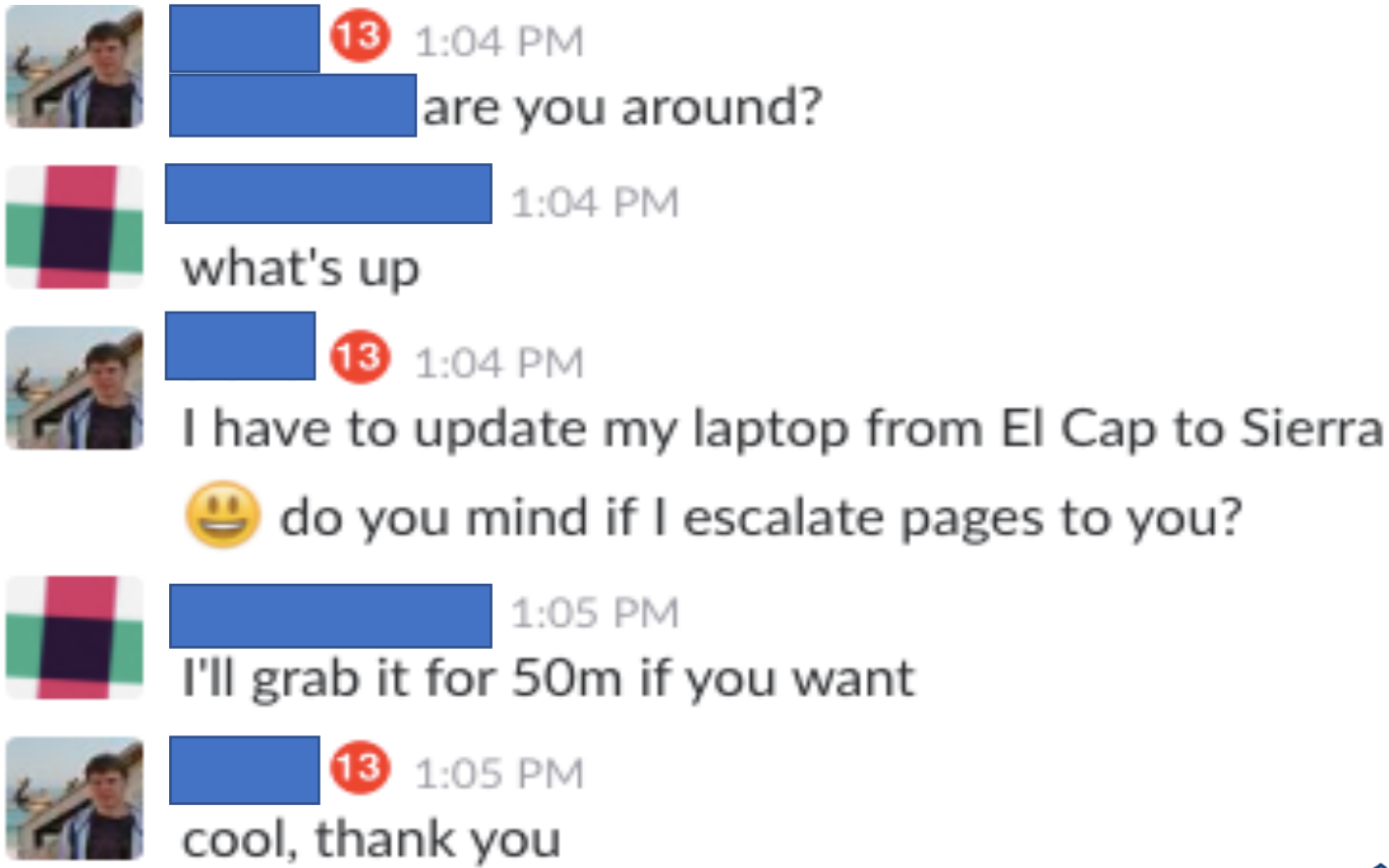


Have we fixed it?



Can we be faster?

- Fast rollback
- Auto-remediation
- Strong 24/7 oncall



Put it all together!





Staying on track

Reliability Measure	Critical System
Participating in Chaos Days	❌
DRTs Completed	✅
Can rapidly rollback	✅
Auto-remediation in place	✅
24/7 on-call rotation in place	✅
AIs Overdue from SEV0s, SEV1s	❌
Not affected by SEV0s	✅



Spotlight: Databases 1 year out

They are in a much better place!

- No high severity incidents since Q3 2016!
- MTTR reduced by 25%!
- Half way through migrating out of a SPOF database!

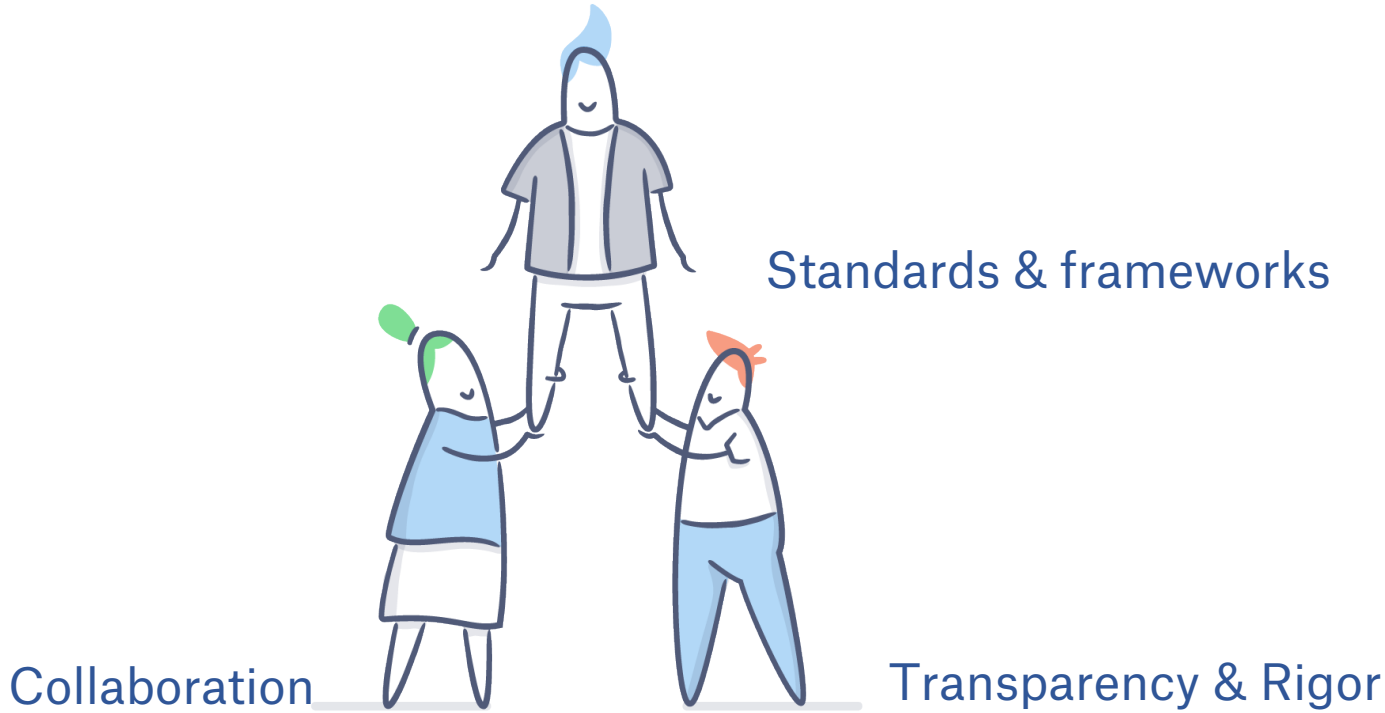


And for the rest of infrastructure?

- ✓ Significant improvement in our availability
- ✓ Better visibility into the systems that matter
- ✓ Better accountability for those systems
- ❑ Limited exposure to the rest of the company



Where are we going?



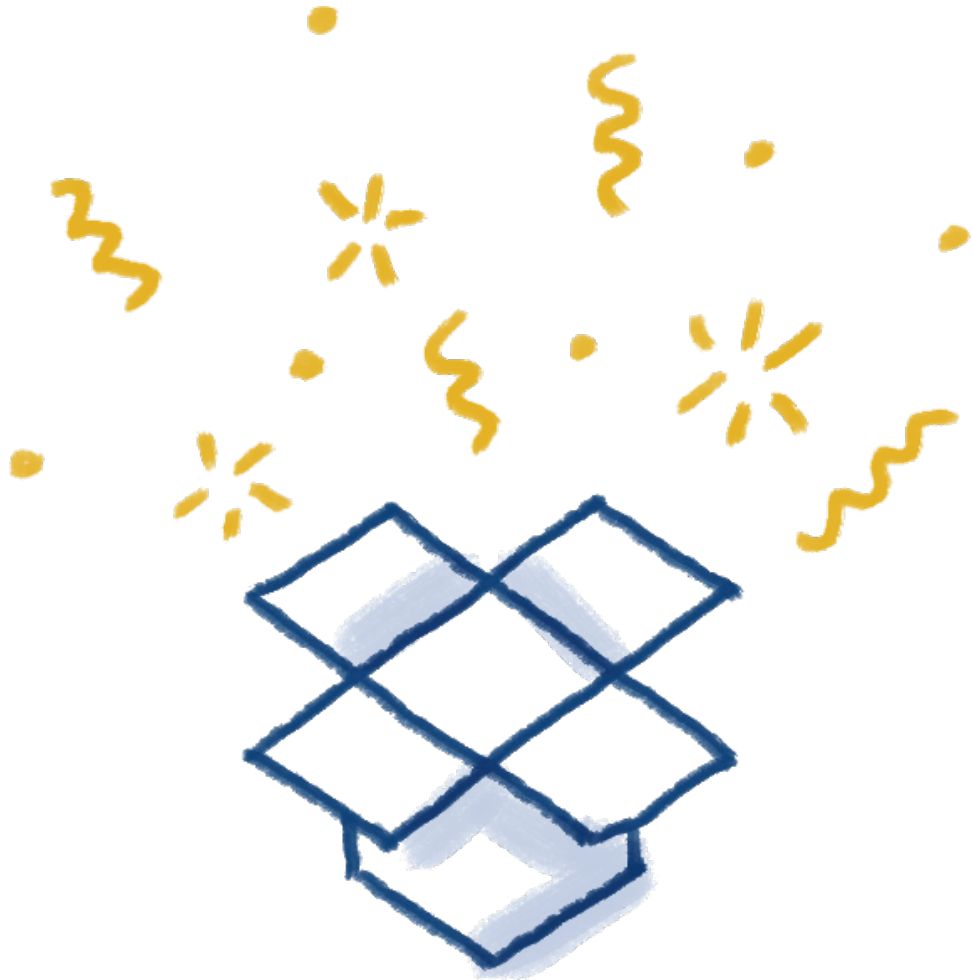
How can you get started?

- Get it going!
- Don’t worry about perfection!
- Learn!



#beworthyoftrust





THANK YOU



Sat Kriya Khalsa
satkriya@dropbox.com