Resilient Response in Complex Systems

John Allspaw
SVP, Tech Ops

Etsy

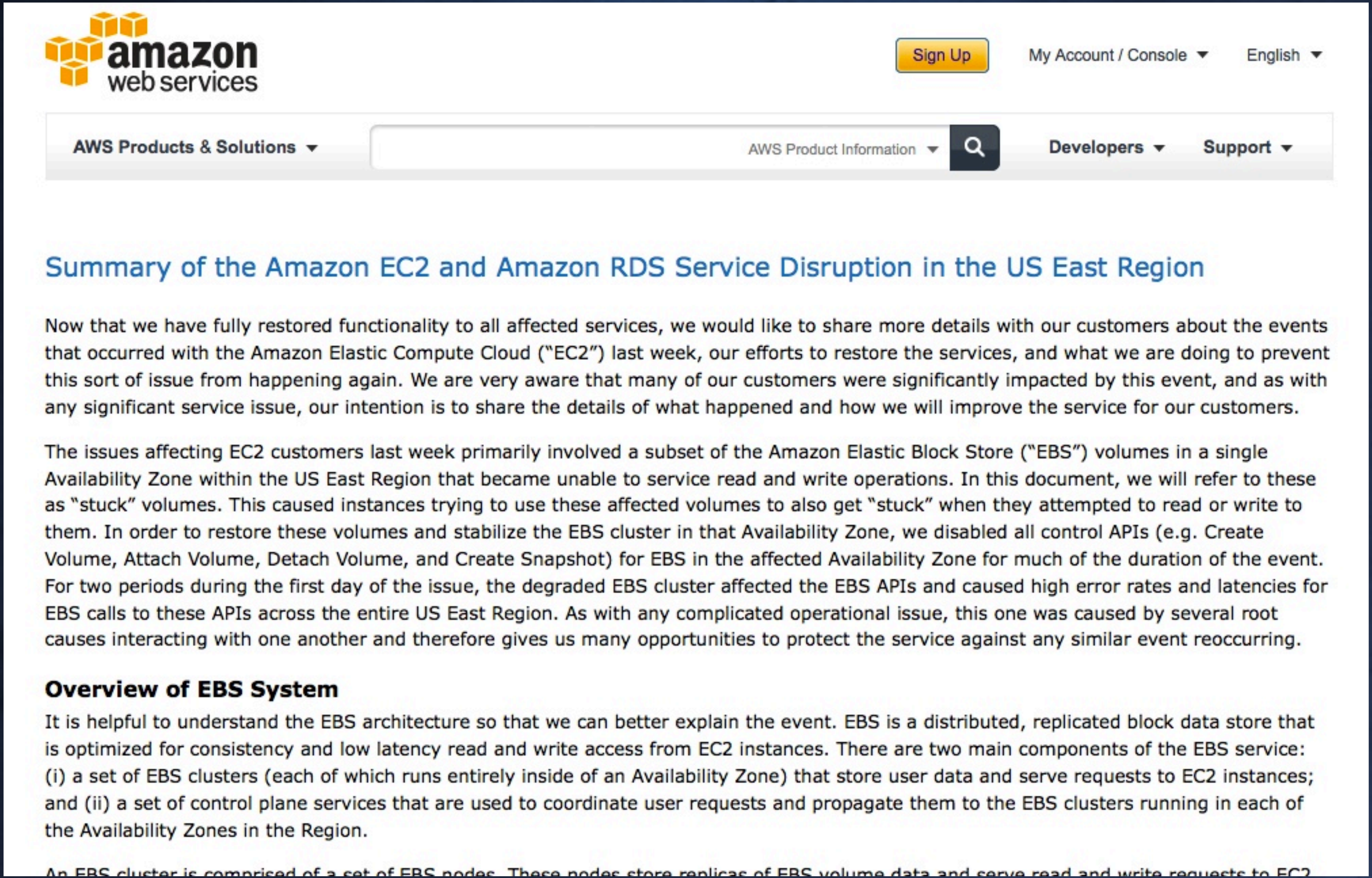
OPERABILITY

PRODUCTION

http://whoownsmyavailability.com

YOU

How important is this?



Summary of the Amazon EC2 and Amazon RDS Service Disruption in the US East Region

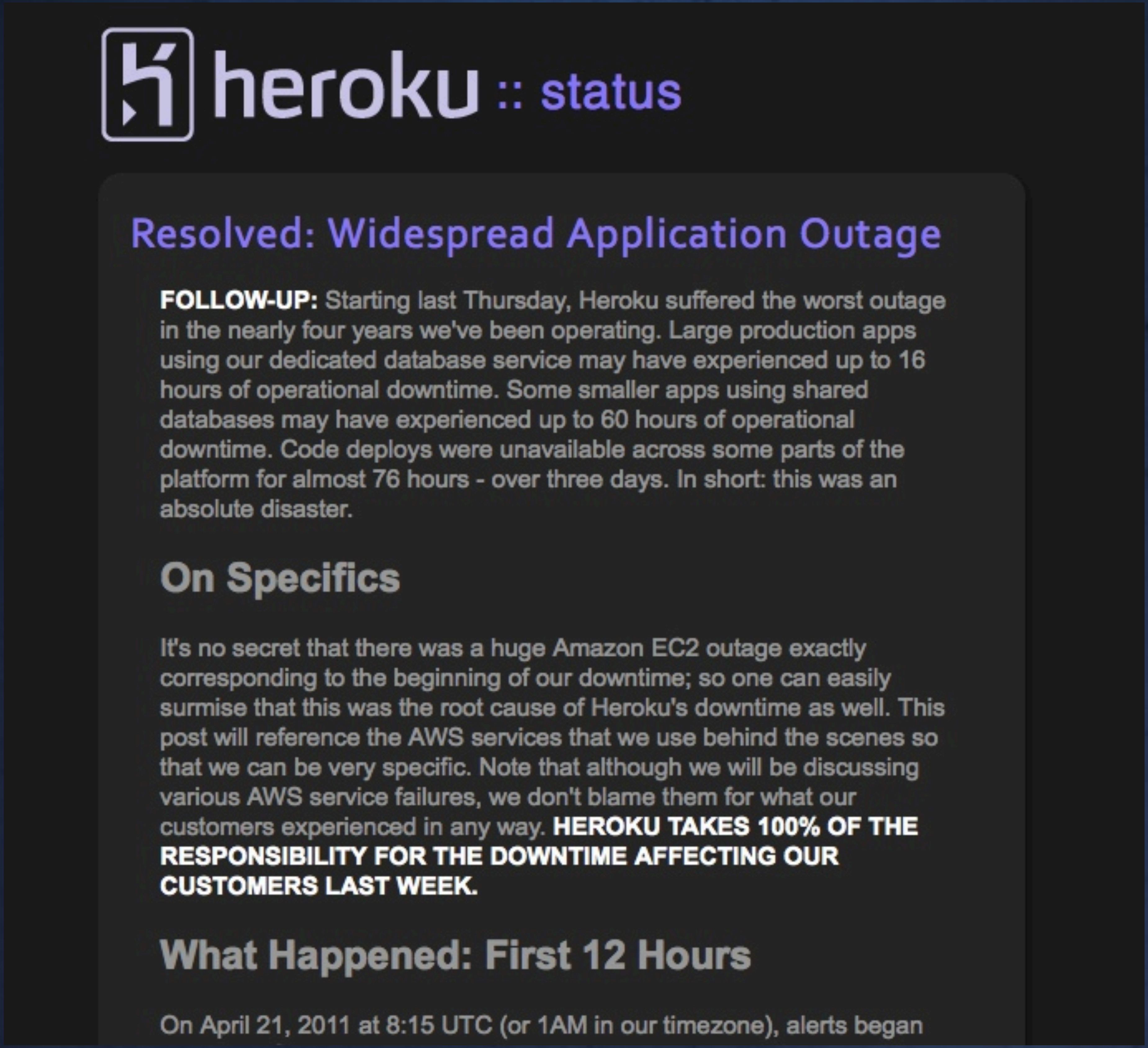
Now that we have fully restored functionality to all affected services, we would like to share more details with our customers about the events that occurred with the Amazon Elastic Compute Cloud ("EC2") last week, our efforts to restore the services, and what we are doing to prevent this sort of issue from happening again. We are very aware that many of our customers were significantly impacted by this event, and as with any significant service issue, our intention is to share the details of what happened and how we will improve the service for our customers.

The issues affecting EC2 customers last week primarily involved a subset of the Amazon Elastic Block Store ("EBS") volumes in a single Availability Zone within the US East Region that became unable to service read and write operations. In this document, we will refer to these as "stuck" volumes. This caused instances trying to use these affected volumes to also get "stuck" when they attempted to read or write to them. In order to restore these volumes and stabilize the EBS cluster in that Availability Zone, we disabled all control APIs (e.g. Create Volume, Attach Volume, Detach Volume, and Create Snapshot) for EBS in the affected Availability Zone for much of the duration of the event. For two periods during the first day of the issue, the degraded EBS cluster affected the EBS APIs and caused high error rates and latencies for EBS calls to these APIs across the entire US East Region. As with any complicated operational issue, this one was caused by several root causes interacting with one another and therefore gives us many opportunities to protect the service against any similar event reoccurring.

Overview of EBS System

It is helpful to understand the EBS architecture so that we can better explain the event. EBS is a distributed, replicated block data store that is optimized for consistency and low latency read and write access from EC2 instances. There are two main components of the EBS service: (i) a set of EBS clusters (each of which runs entirely inside of an Availability Zone) that store user data and serve requests to EC2 instances; and (ii) a set of control plane services that are used to coordinate user requests and propagate them to the EBS clusters running in each of the Availability Zones in the Region.

An EBS cluster is comprised of a set of EBS nodes. These nodes store replicas of EBS volume data and serve read and write requests to EC2



Resolved: Widespread Application Outage

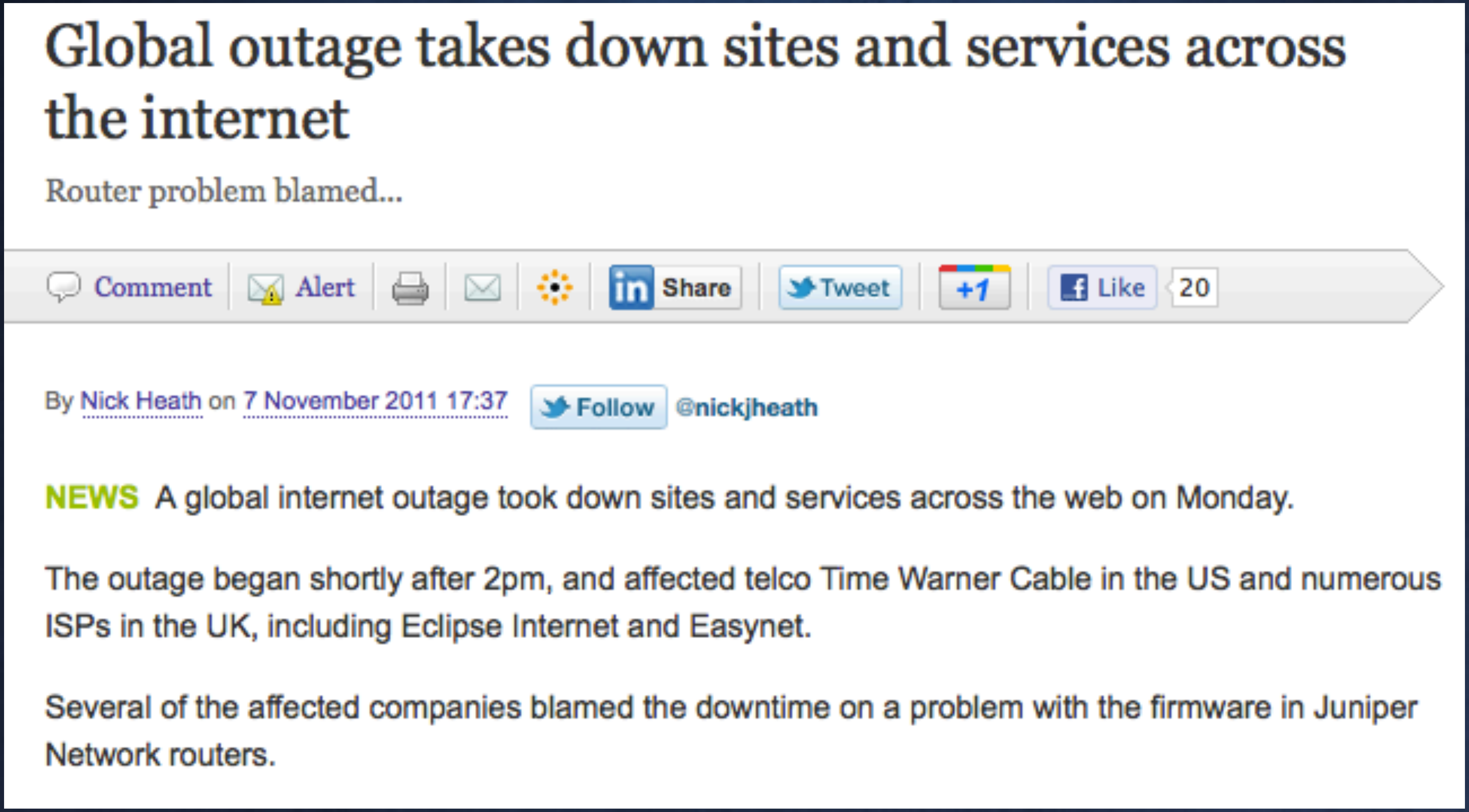
FOLLOW-UP: Starting last Thursday, Heroku suffered the worst outage in the nearly four years we've been operating. Large production apps using our dedicated database service may have experienced up to 16 hours of operational downtime. Some smaller apps using shared databases may have experienced up to 60 hours of operational downtime. Code deploys were unavailable across some parts of the platform for almost 76 hours - over three days. In short: this was an absolute disaster.

On Specifics

It's no secret that there was a huge Amazon EC2 outage exactly corresponding to the beginning of our downtime; so one can easily surmise that this was the root cause of Heroku's downtime as well. This post will reference the AWS services that we use behind the scenes so that we can be very specific. Note that although we will be discussing various AWS service failures, we don't blame them for what our customers experienced in any way. **HEROKU TAKES 100% OF THE RESPONSIBILITY FOR THE DOWNTIME AFFECTING OUR CUSTOMERS LAST WEEK.**

What Happened: First 12 Hours

On April 21, 2011 at 8:15 UTC (or 1AM in our timezone), alerts began



Global outage takes down sites and services across the internet

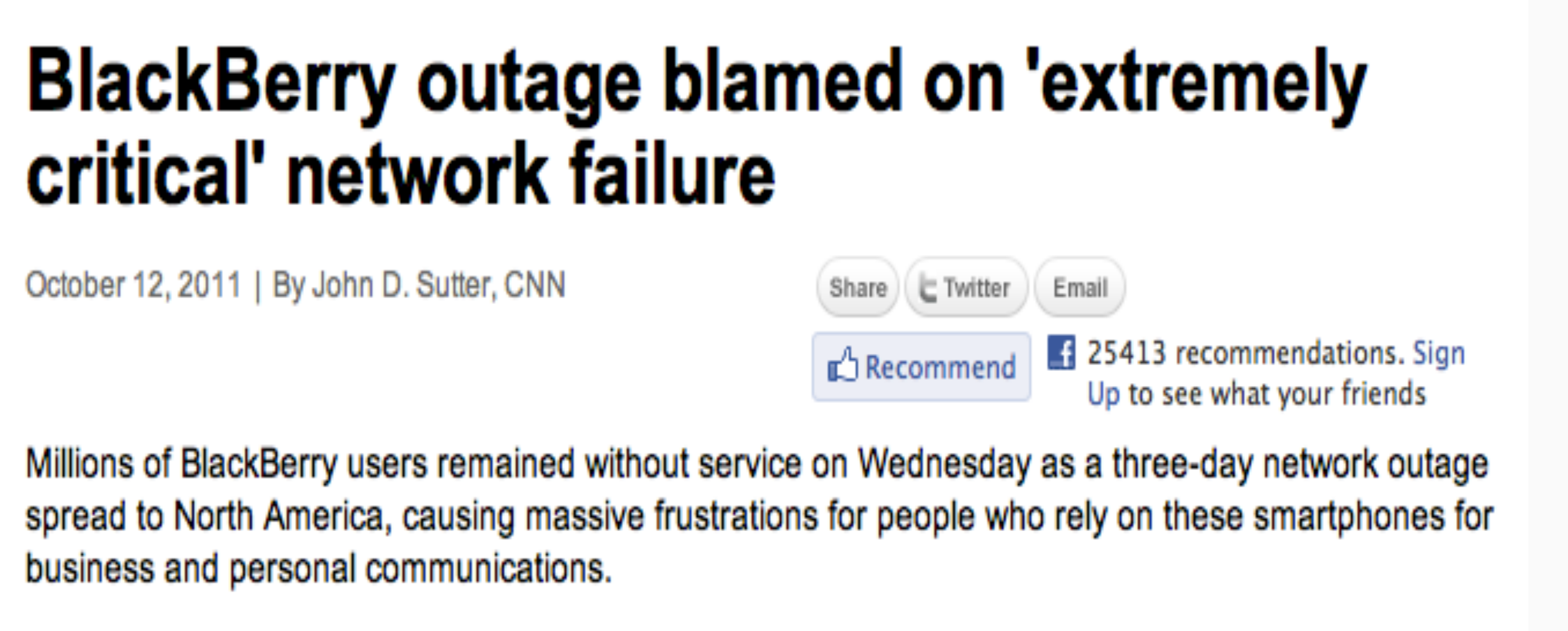
Router problem blamed...

By Nick Heath on 7 November 2011 17:37 @nickjheath

NEWS A global internet outage took down sites and services across the web on Monday.

The outage began shortly after 2pm, and affected telco Time Warner Cable in the US and numerous ISPs in the UK, including Eclipse Internet and Easynet.

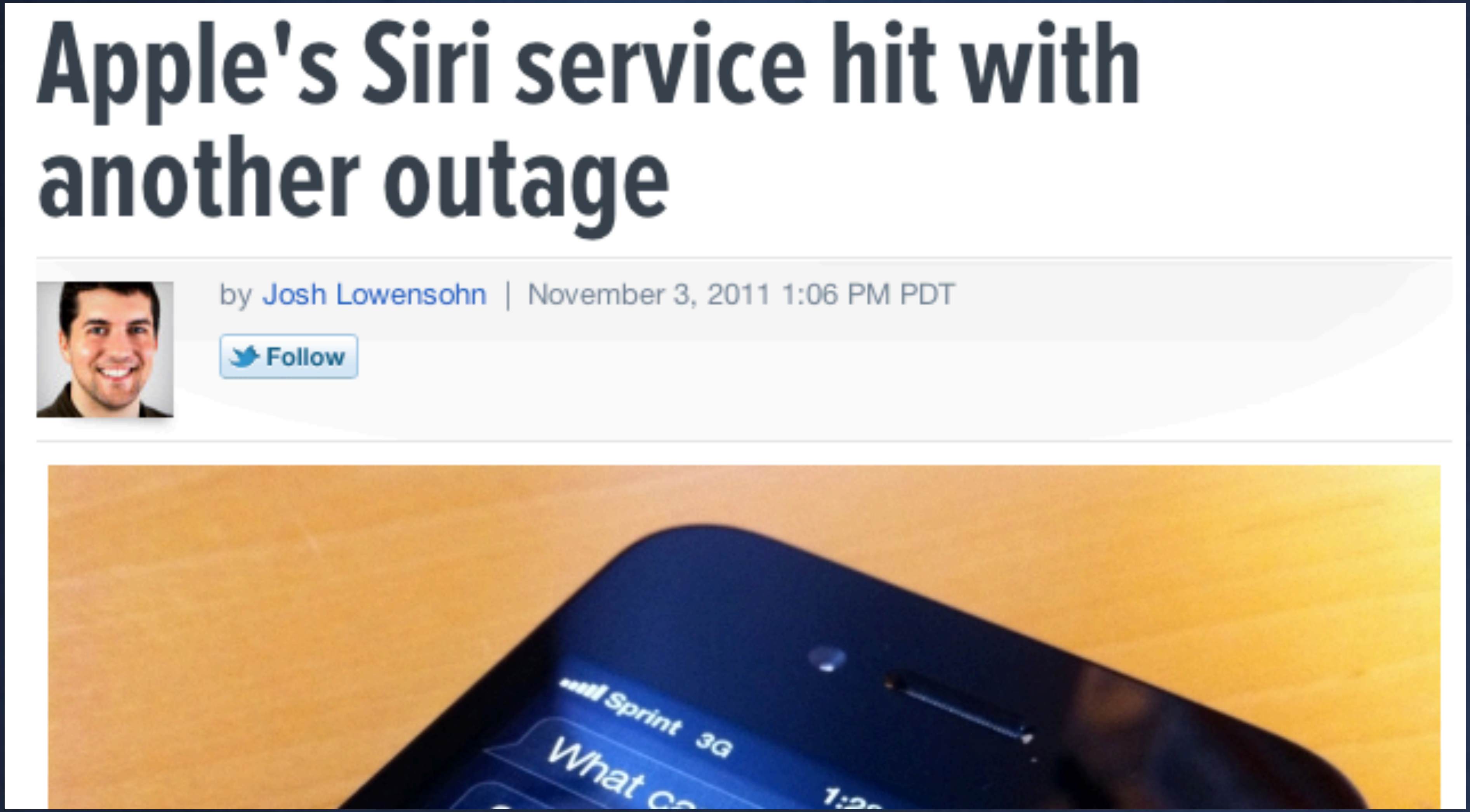
Several of the affected companies blamed the downtime on a problem with the firmware in Juniper Network routers.



BlackBerry outage blamed on 'extremely critical' network failure

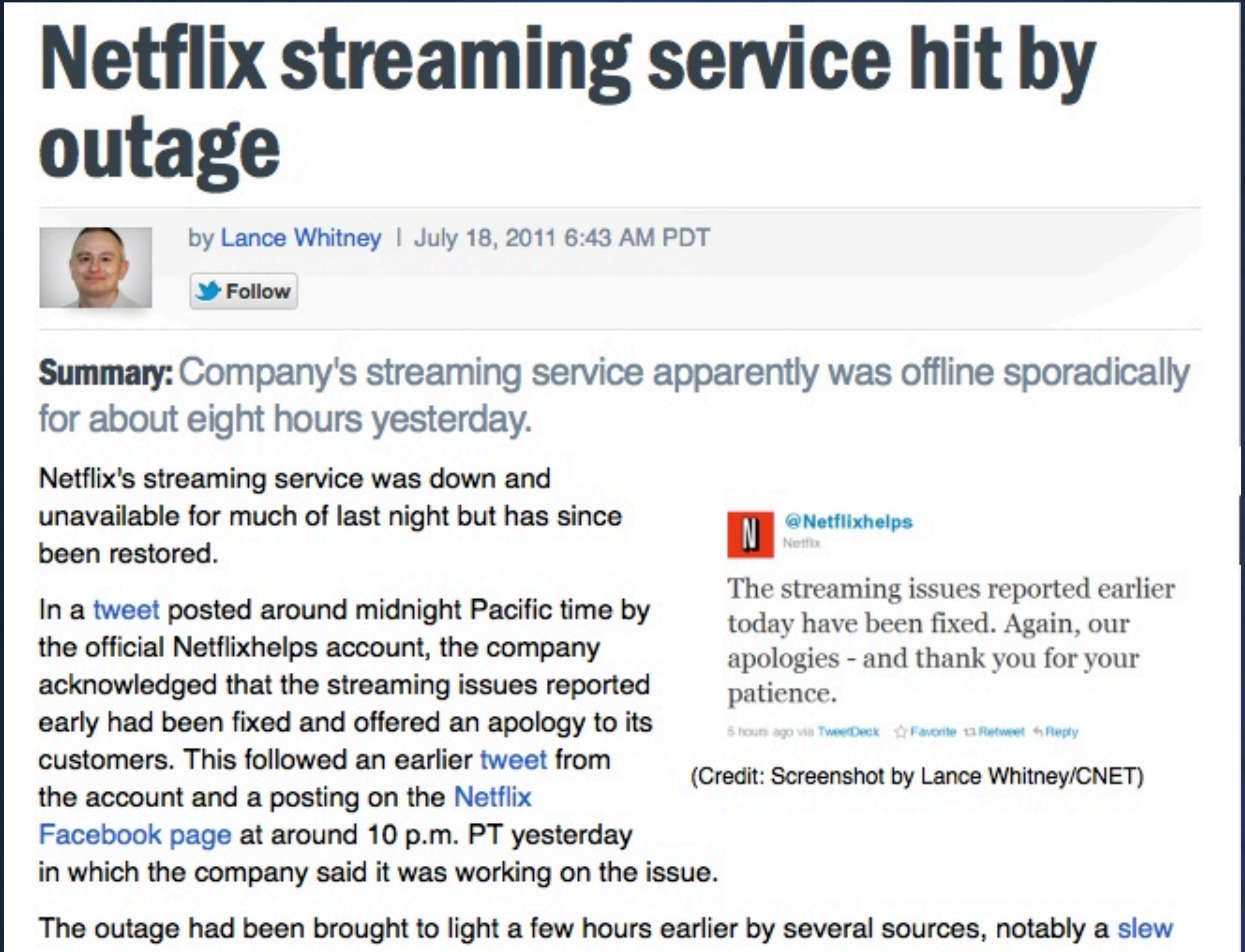
October 12, 2011 | By John D. Sutter, CNN

Millions of BlackBerry users remained without service on Wednesday as a three-day network outage spread to North America, causing massive frustrations for people who rely on these smartphones for business and personal communications.



Apple's Siri service hit with another outage

by Josh Lowensohn | November 3, 2011 1:06 PM PDT



Netflix streaming service hit by outage

by Lance Whitney | July 18, 2011 6:43 AM PDT

Summary: Company's streaming service apparently was offline sporadically for about eight hours yesterday.

Netflix's streaming service was down and unavailable for much of last night but has since been restored.

In a tweet posted around midnight Pacific time by the official Netflixhelps account, the company acknowledged that the streaming issues reported early had been fixed and offered an apology to its customers. This followed an earlier tweet from the account and a posting on the Netflix Facebook page at around 10 p.m. PT yesterday in which the company said it was working on the issue.

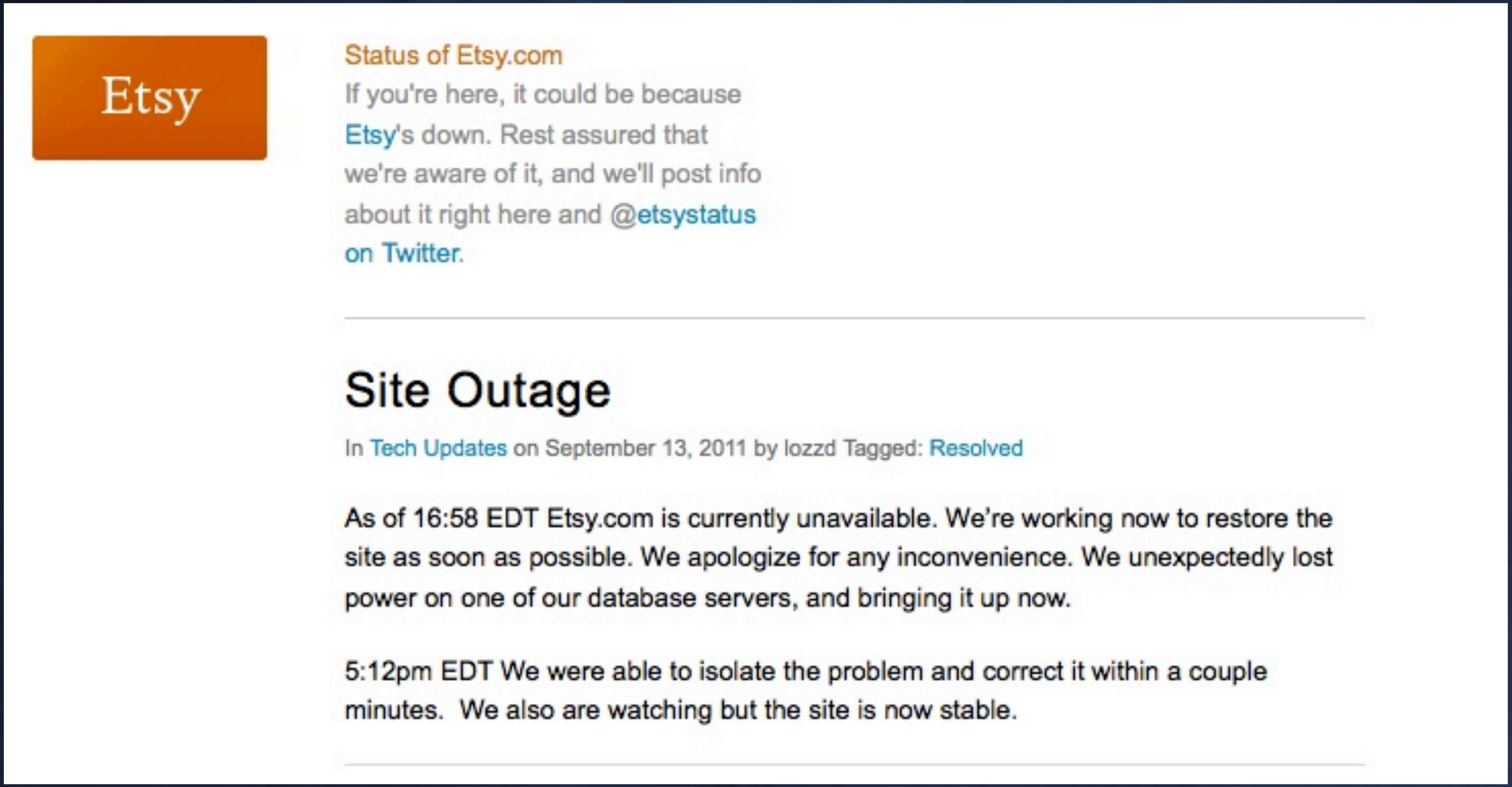
@Netflixhelps
Netflix

The streaming issues reported earlier today have been fixed. Again, our apologies - and thank you for your patience.

5 hours ago via TweetDeck ☆ Favorite ⇄ Retweet ↰ Reply

(Credit: Screenshot by Lance Whitney/CNET)

The outage had been brought to light a few hours earlier by several sources, notably a slew



Status of Etsy.com

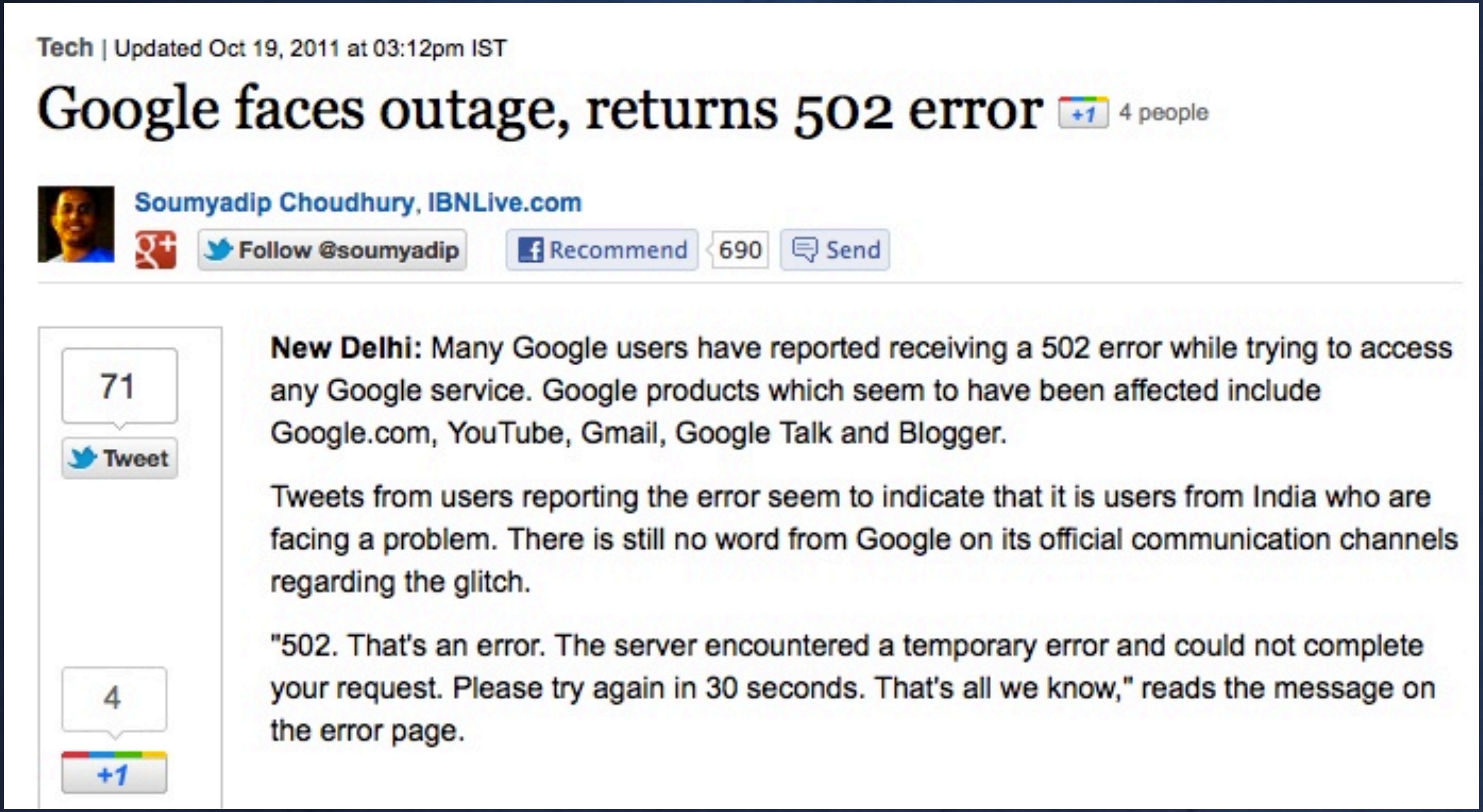
If you're here, it could be because Etsy's down. Rest assured that we're aware of it, and we'll post info about it right here and @etsystatus on Twitter.

Site Outage

In Tech Updates on September 13, 2011 by lozzd Tagged: Resolved

As of 16:58 EDT Etsy.com is currently unavailable. We're working now to restore the site as soon as possible. We apologize for any inconvenience. We unexpectedly lost power on one of our database servers, and bringing it up now.

5:12pm EDT We were able to isolate the problem and correct it within a couple minutes. We also are watching but the site is now stable.



Tech | Updated Oct 19, 2011 at 03:12pm IST

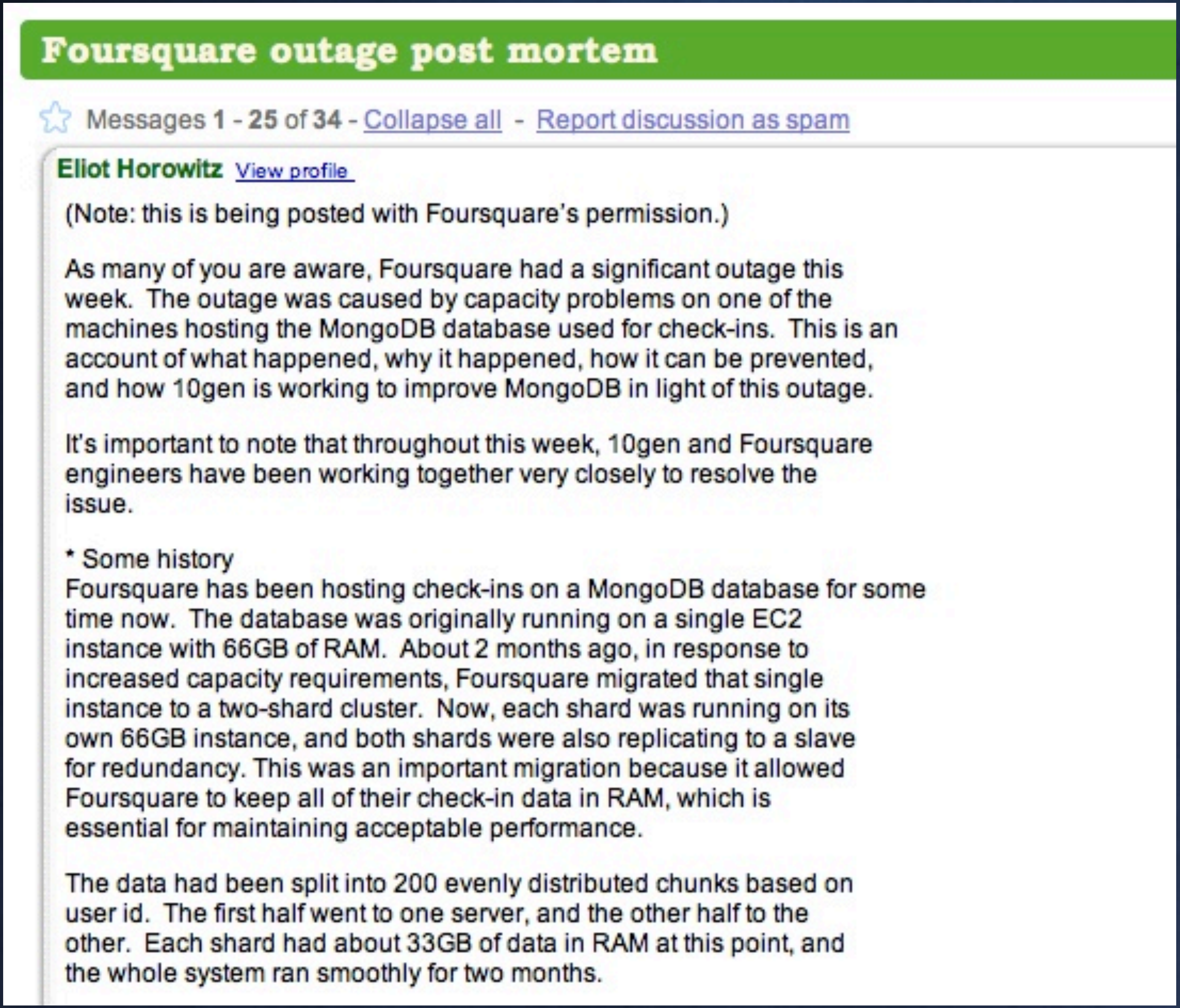
Google faces outage, returns 502 error

Soumyadip Choudhury, IBNLive.com

New Delhi: Many Google users have reported receiving a 502 error while trying to access any Google service. Google products which seem to have been affected include Google.com, YouTube, Gmail, Google Talk and Blogger.

Tweets from users reporting the error seem to indicate that it is users from India who are facing a problem. There is still no word from Google on its official communication channels regarding the glitch.

"502. That's an error. The server encountered a temporary error and could not complete your request. Please try again in 30 seconds. That's all we know," reads the message on the error page.



Foursquare outage post mortem

☆ Messages 1 - 25 of 34 - Collapse all - Report discussion as spam

Eliot Horowitz

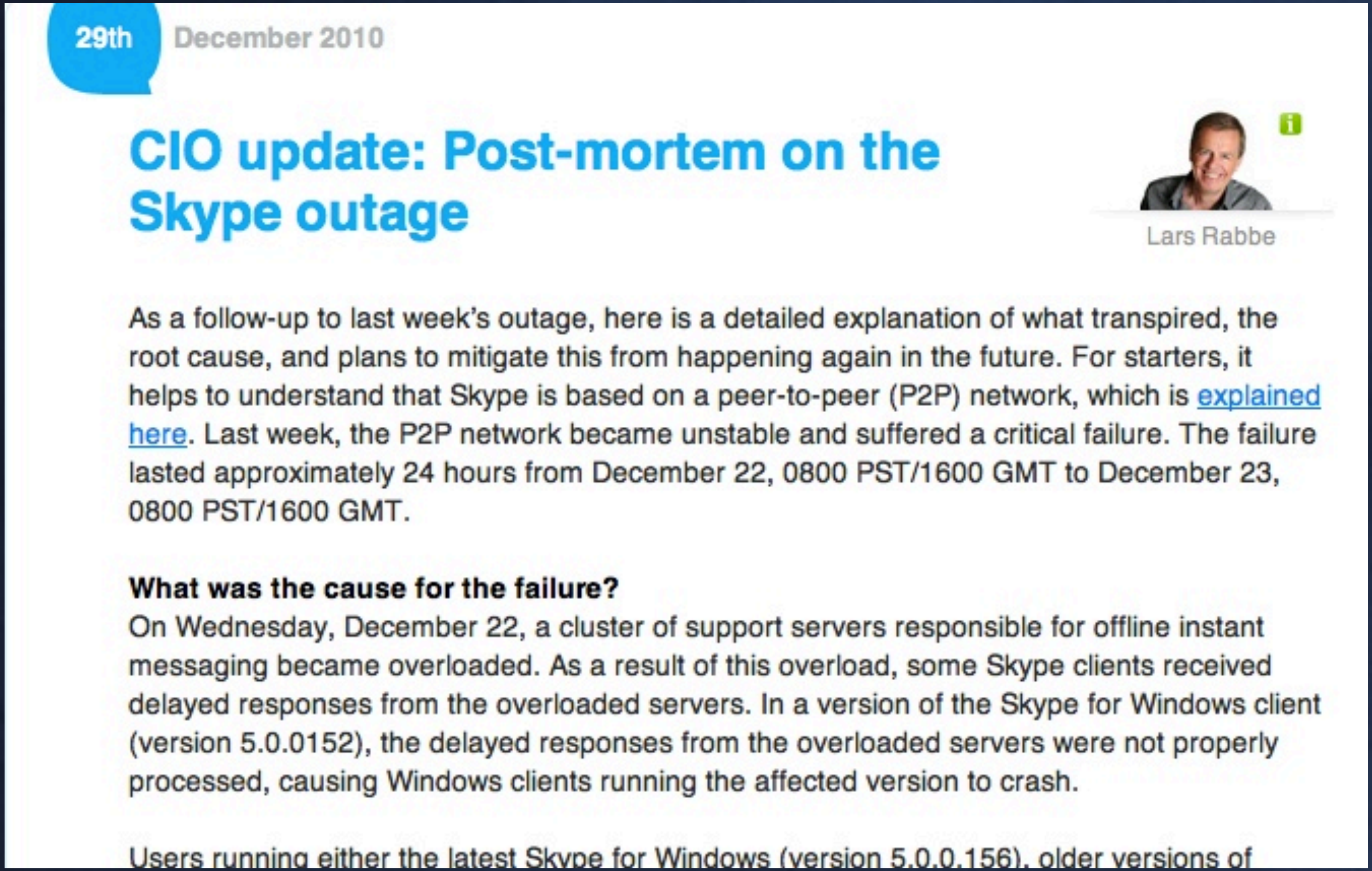
(Note: this is being posted with Foursquare's permission.)

As many of you are aware, Foursquare had a significant outage this week. The outage was caused by capacity problems on one of the machines hosting the MongoDB database used for check-ins. This is an account of what happened, why it happened, how it can be prevented, and how 10gen is working to improve MongoDB in light of this outage.

It's important to note that throughout this week, 10gen and Foursquare engineers have been working together very closely to resolve the issue.

* Some history
Foursquare has been hosting check-ins on a MongoDB database for some time now. The database was originally running on a single EC2 instance with 66GB of RAM. About 2 months ago, in response to increased capacity requirements, Foursquare migrated that single instance to a two-shard cluster. Now, each shard was running on its own 66GB instance, and both shards were also replicating to a slave for redundancy. This was an important migration because it allowed Foursquare to keep all of their check-in data in RAM, which is essential for maintaining acceptable performance.

The data had been split into 200 evenly distributed chunks based on user id. The first half went to one server, and the other half to the other. Each shard had about 33GB of data in RAM at this point, and the whole system ran smoothly for two months.



29th December 2010

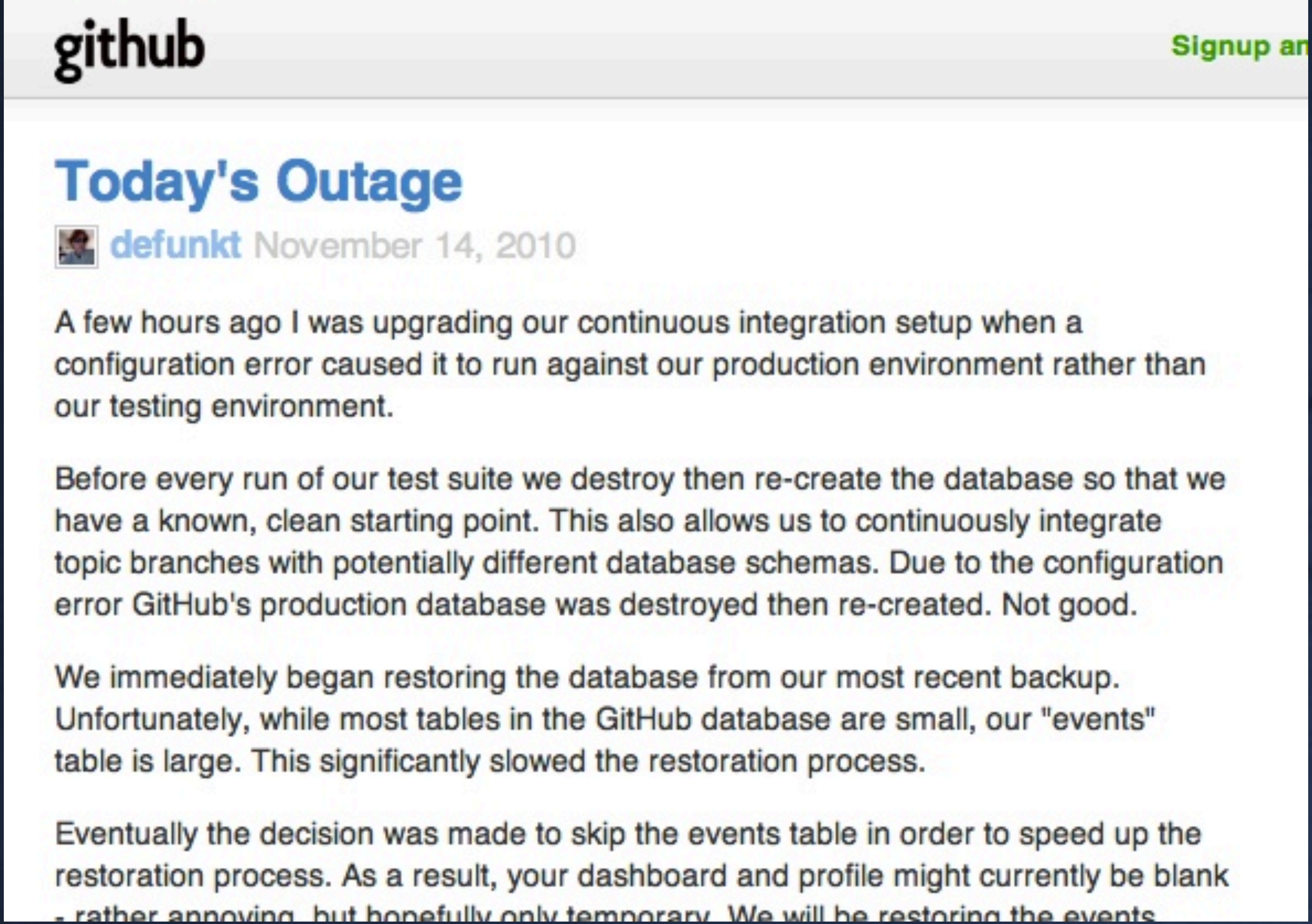
CIO update: Post-mortem on the Skype outage

Lars Rabbe

As a follow-up to last week's outage, here is a detailed explanation of what transpired, the root cause, and plans to mitigate this from happening again in the future. For starters, it helps to understand that Skype is based on a peer-to-peer (P2P) network, which is explained here. Last week, the P2P network became unstable and suffered a critical failure. The failure lasted approximately 24 hours from December 22, 0800 PST/1600 GMT to December 23, 0800 PST/1600 GMT.

What was the cause for the failure?
On Wednesday, December 22, a cluster of support servers responsible for offline instant messaging became overloaded. As a result of this overload, some Skype clients received delayed responses from the overloaded servers. In a version of the Skype for Windows client (version 5.0.0152), the delayed responses from the overloaded servers were not properly processed, causing Windows clients running the affected version to crash.

Users running either the latest Skype for Windows (version 5.0.0.156), older versions of



Today's Outage

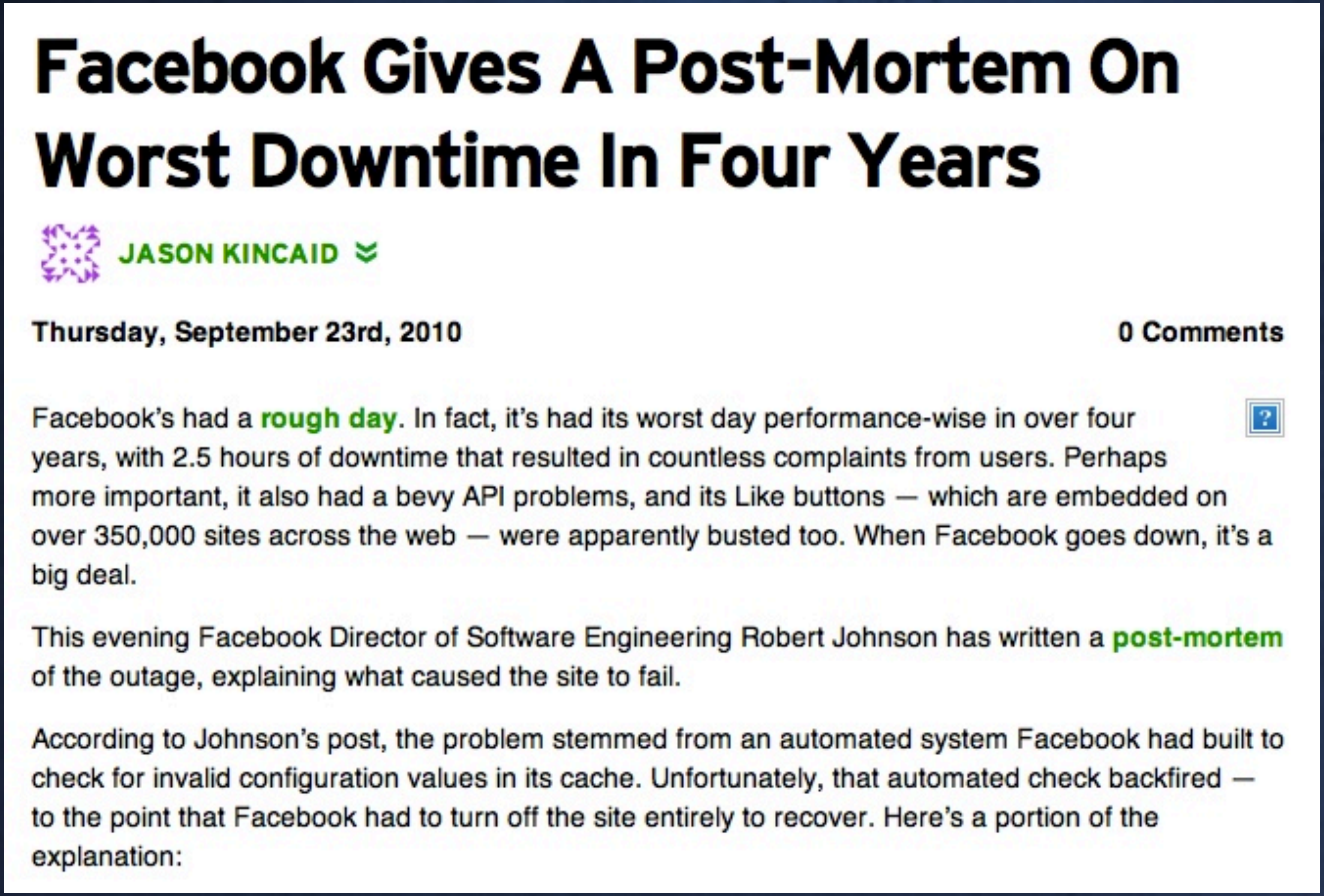
defunkt November 14, 2010

A few hours ago I was upgrading our continuous integration setup when a configuration error caused it to run against our production environment rather than our testing environment.

Before every run of our test suite we destroy then re-create the database so that we have a known, clean starting point. This also allows us to continuously integrate topic branches with potentially different database schemas. Due to the configuration error GitHub's production database was destroyed then re-created. Not good.

We immediately began restoring the database from our most recent backup. Unfortunately, while most tables in the GitHub database are small, our "events" table is large. This significantly slowed the restoration process.

Eventually the decision was made to skip the events table in order to speed up the restoration process. As a result, your dashboard and profile might currently be blank - rather annoying, but hopefully only temporary. We will be restoring the events



Facebook Gives A Post-Mortem On Worst Downtime In Four Years

JASON KINCAID

Thursday, September 23rd, 2010 **0 Comments**

Facebook's had a **rough day**. In fact, it's had its worst day performance-wise in over four years, with 2.5 hours of downtime that resulted in countless complaints from users. Perhaps more important, it also had a bevy API problems, and its Like buttons — which are embedded on over 350,000 sites across the web — were apparently busted too. When Facebook goes down, it's a big deal.

This evening Facebook Director of Software Engineering Robert Johnson has written a **post-mortem** of the outage, explaining what caused the site to fail.

According to Johnson's post, the problem stemmed from an automated system Facebook had built to check for invalid configuration values in its cache. Unfortunately, that automated check backfired — to the point that Facebook had to turn off the site entirely to recover. Here's a portion of the explanation:

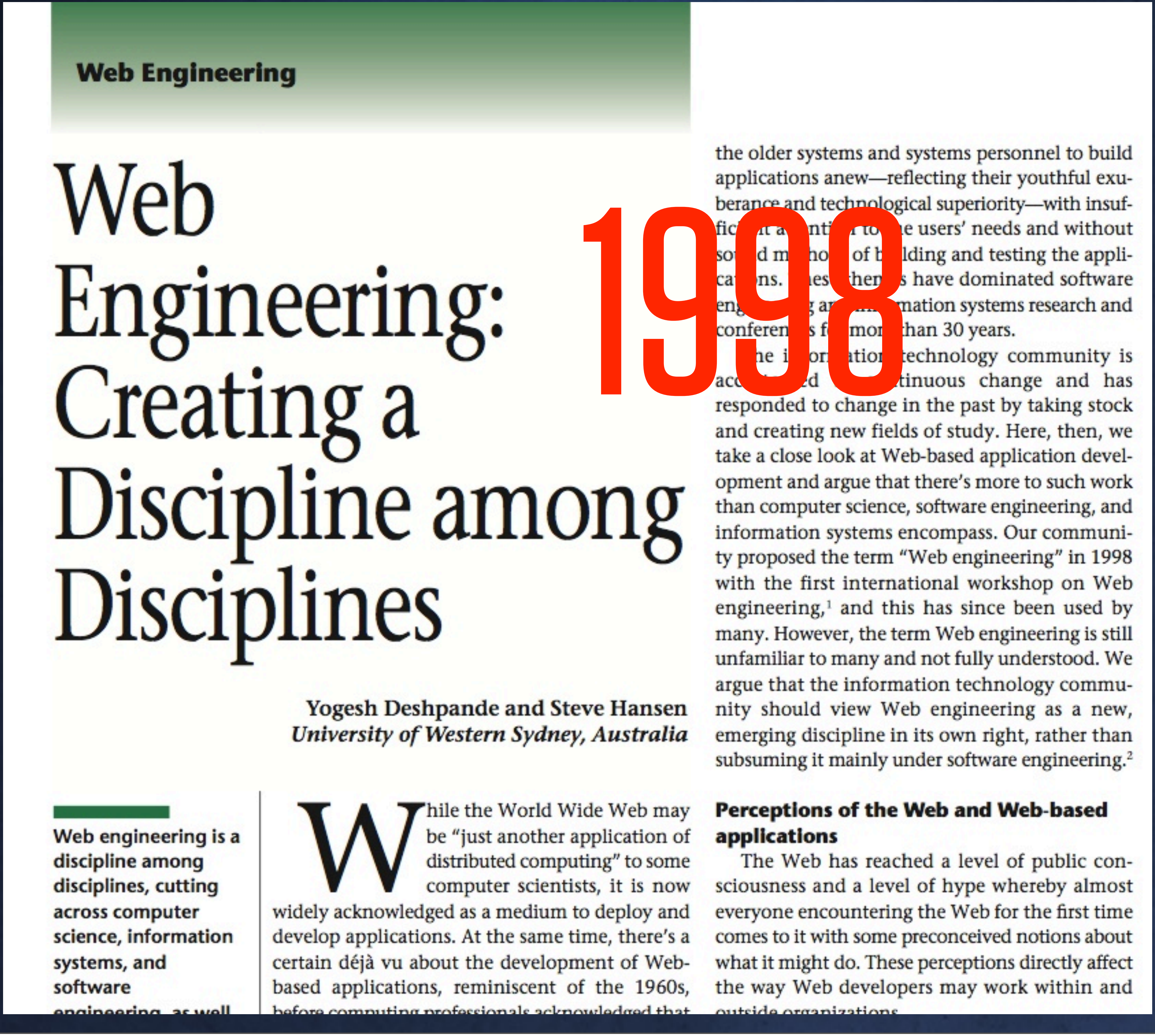
How important is this?

How Can This Happen?

1636 655
1636 655

Complex Systems

- *Cascading Failures*
- *Difficult to determine boundaries*
- *Complex systems may be open*
- *Complex systems may have a memory*
- *Complex systems may be nested*
- *Dynamic network of multiplicity*
- *May produce emergent phenomena*
- *Relationships are non-linear*
- *Relationships contain feedback loops*



1998

Web Engineering

Web Engineering: Creating a Discipline among Disciplines

Yogesh Deshpande and Steve Hansen
University of Western Sydney, Australia

Web engineering is a discipline among disciplines, cutting across computer science, information systems, and software engineering, as well

While the World Wide Web may be "just another application of distributed computing" to some computer scientists, it is now widely acknowledged as a medium to deploy and develop applications. At the same time, there's a certain déjà vu about the development of Web-based applications, reminiscent of the 1960s, before computing professionals acknowledged that the older systems and systems personnel to build applications anew—reflecting their youthful exuberance and technological superiority—with insufficient attention to the users' needs and without sound methods of building and testing the applications. These themes have dominated software engineering and information systems research and conferences for more than 30 years.

The information technology community is accustomed to continuous change and has responded to change in the past by taking stock and creating new fields of study. Here, then, we take a close look at Web-based application development and argue that there's more to such work than computer science, software engineering, and information systems encompass. Our community proposed the term "Web engineering" in 1998 with the first international workshop on Web engineering,[1] and this has since been used by many. However, the term Web engineering is still unfamiliar to many and not fully understood. We argue that the information technology community should view Web engineering as a new, emerging discipline in its own right, rather than subsuming it mainly under software engineering.[2]

Perceptions of the Web and Web-based applications

The Web has reached a level of public consciousness and a level of hype whereby almost everyone encountering the Web for the first time comes to it with some preconceived notions about what it might do. These perceptions directly affect the way Web developers may work within and outside organizations.

How Can This Happen?

It does happen.

And it will again.

And again.

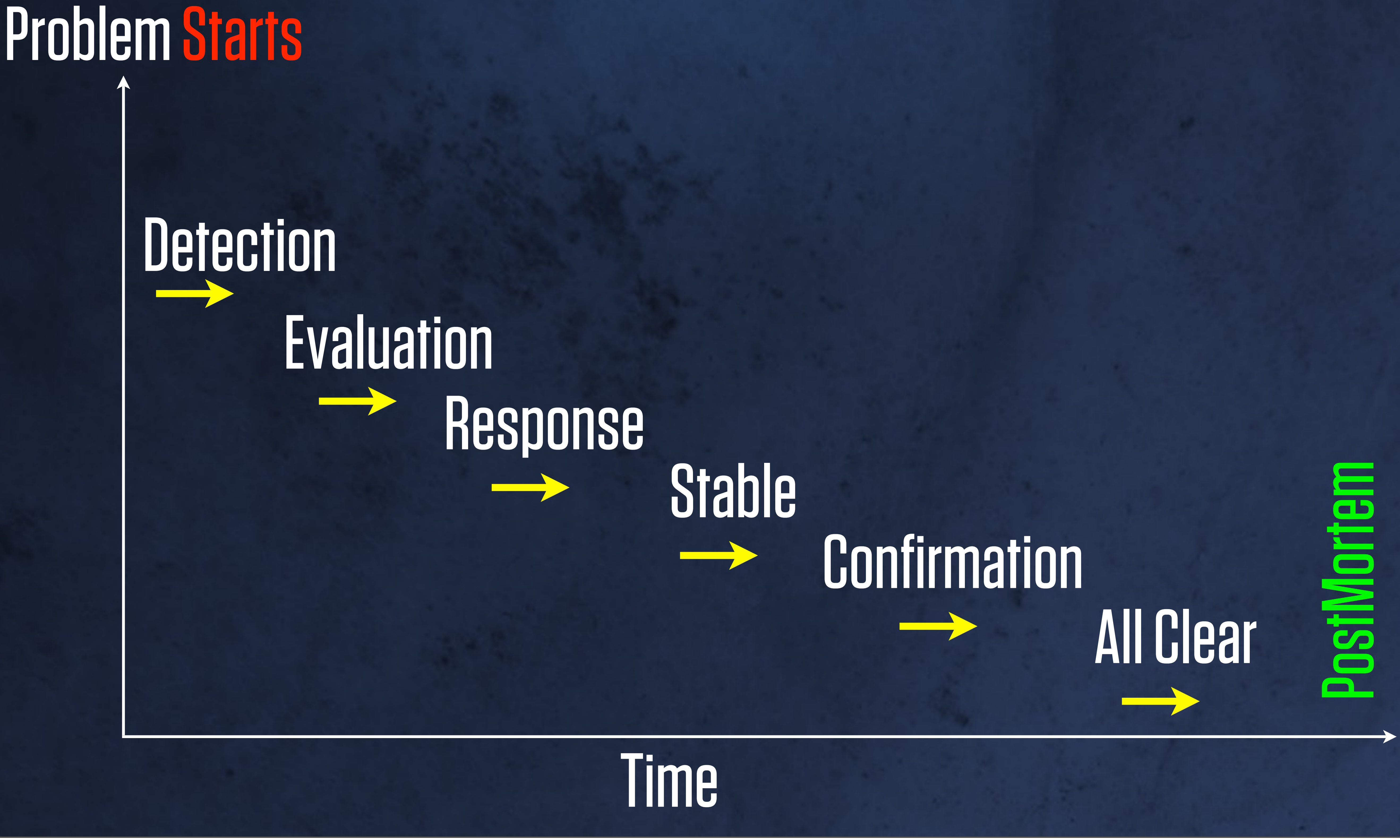
Optimization

~~MTBF~~

MTTR

http://www.flickr.com/photos/sparktography/75499095/

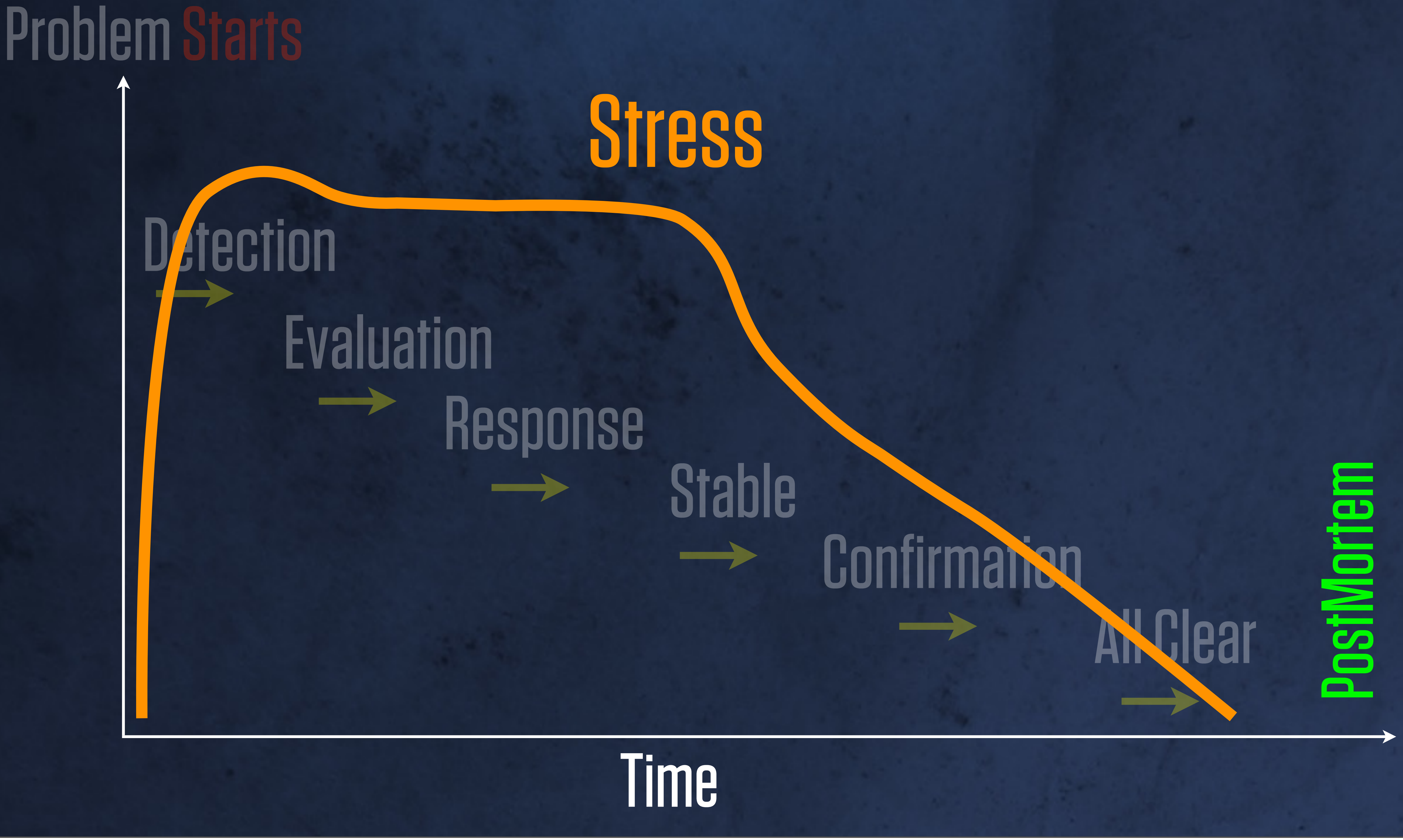
How does team troubleshooting happen?



Problem Starts

Detection → Evaluation → Response → Stable → Confirmation → All Clear → PostMortem

Time


Problem Starts
Stress
Detection
Evaluation
Response
Stable
Confirmation
All Clear
PostMortem
Time

Forced beyond learned roles

Actions whose consequences are both important and difficult to see

Cognitively and perceptively noisy

Coordinative load increases exponentially

So What Can We Do?

We Learn From Others

Characteristics of response to escalating scenarios

Characteristics of response to escalating scenarios

...tend to neglect how processes develop within time (awareness of rates) versus assessing how things are in the moment

"On the Difficulties People Have in Dealing With Complexity" Dietrich Doerner, 1980

Characteristics of response to escalating scenarios

...have difficulty in dealing with exponential developments (hard to imagine how fast something can change, or accelerate)

"On the Difficulties People Have in Dealing With Complexity" Dietrich Doerner, 1980

Characteristics of response to escalating scenarios

...inclined to think in causal series, instead of causal nets.

A therefore B,

instead of

A, therefore B and C (therefore D and E), etc.

“On the Difficulties People Have in Dealing With Complexity” Dietrich Doerner, 1980

Pitfalls

Thematic Vagabonding

Pitfalls

Goal Fixation

(encystment)

Pitfalls

Refusal to make decisions

Heroism

Non-communicating lone wolf-isms

Distraction

Irrelevant noise in comm channels

Jens Rasmussen, 1983

Senior Member, IEEE

"Skills, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models"

IEEE Transactions On Systems, Man, and Cybernetics, May 1983

SKILL - BASED

Simple, routine

RULE - BASED

Knowable, but unfamiliar

KNOWLEDGE - BASED

WTF IS GOING ON?

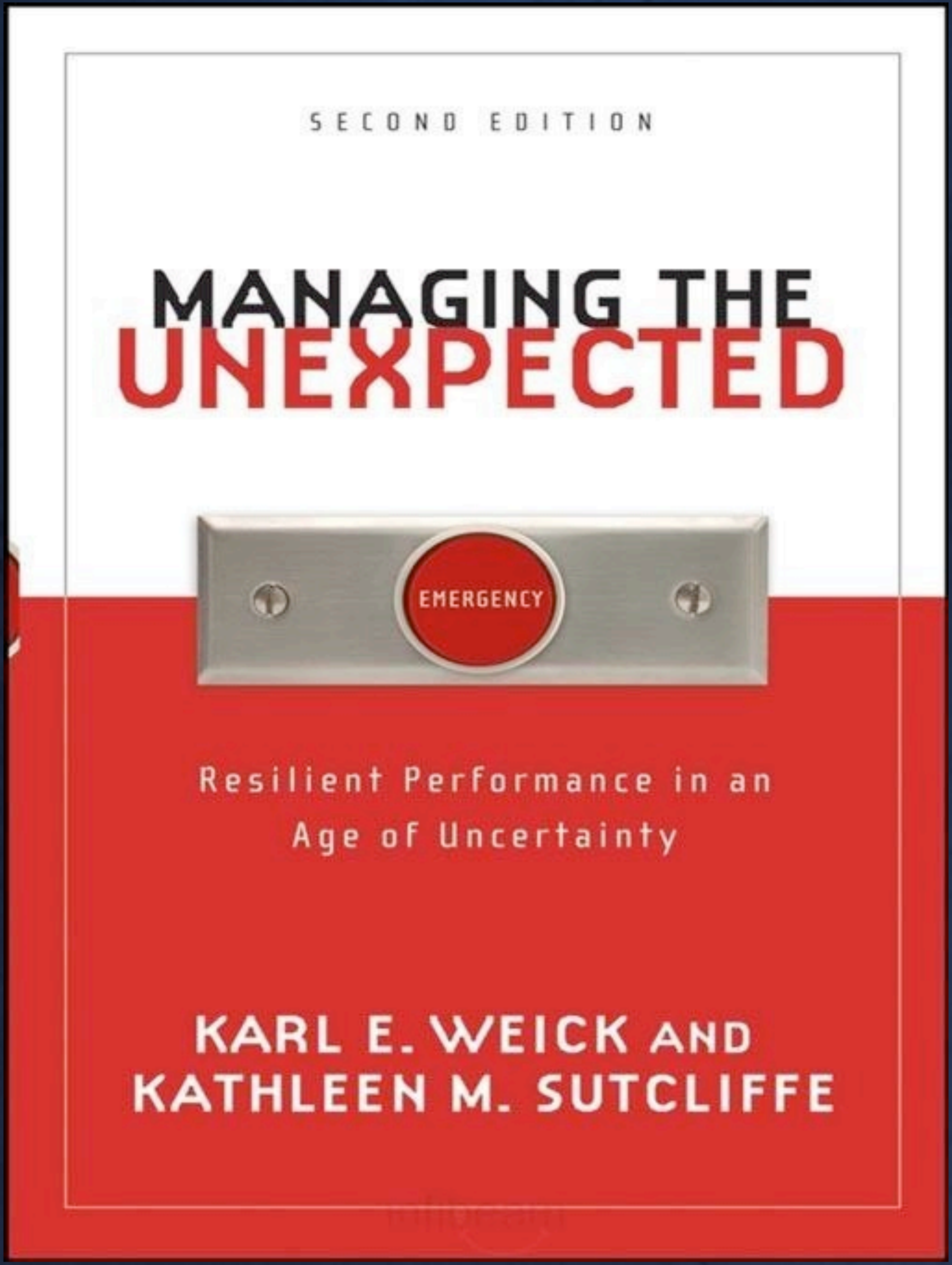
(Reason, 1990)

Team Dynamics

High Reliability Organizations

- Air Traffic Control
- Naval Air Operations At Sea
- Electrical Power Systems
- Etc.

- Complex Socio-Technical systems
- Efficiency <-> Thoroughness
- Time/Resource Constrained
- Engineering-driven



SECOND EDITION

MANAGING THE UNEXPECTED

Resilient Performance in an Age of Uncertainty

KARL E. WEICK AND KATHLEEN M. SUTCLIFFE

“The Self-Designing High-Reliability Organization: Aircraft Carrier Flight Operations at Sea”

Rochlin, La Porte, and Roberts. Naval War College Review 1987

Close interdependence between groups

Close reciprocal coordination and information sharing, resulting in overlapping knowledge

High redundancy: multiple people observing the same event and sharing information

Broad definition of who belongs to the team.

Teammates are included in the communication loops rather than excluded.

Lots of error correction.

High levels of situation comprehension: maintain constant awareness of the possibility of accidents.

High levels of interpersonal skills

Maintenance of detailed records of past incidents that are closely examined with a view to learning from them.

Patterns of authority are changed to meet the demands of the events: organizational flexibility.

The reporting of errors and faults is rewarded, not punished.

So What Else Can We Do?

We Drill

We GameDay

We Learn To Improvise

IMPROVISATION

IMPROVISATION

We Learn From Our Mistakes

Postmortems

- Full timelines: What happened, when
- Review in public, everyone invited
- Search for “second stories” instead of “human error”
- Cultivating a blameless environment
- Giving requisite authority to individuals to improve things

Qualifying Response

High signal:noise in comm channels?

Troubleshooting fatigue?

Troubleshooting handoff?

All tools on-hand?

Improvised tooling or solutions?

Metrics visibility?

Collaborative and skillful communication?

Remediation

Mature Role of Automation

"Ironies of Automation" - Lisanne Bainbridge

http://www.bainbrdg.demon.co.uk/Papers/Ironies.html

Mature Role of Automation

- Moves humans from manual operator to supervisor
- Extends and augments human abilities, doesn’t replace it
- Doesn’t remove “human error”
- Are brittle
- Recognize that there is always discretionary space for humans
- Recognizes the Law of Stretched Systems

Law of Stretched Systems

"Every system is stretched to operate at its capacity; as soon as there is some improvement, for example, in the form of new technology, it will be exploited to achieve a new intensity and tempo of activity"

D. Woods, E. Hollnagel, "Joint Cognitive Systems: Patterns" 2006

We Share Near-Miss Events

Near Misses

Hey everybody -

Don’t be like me. I tried to X, but that wasn’t a good idea.

It almost exploded everyone.

So, don’t do: *(details about X)*

Love,
Joe

Near Misses

- Can act like “vaccines” - help system safety without actually hurting anything
- Happen more often, so provide more data on latent failures
- Powerful reminder of hazards, and slows down the process of forgetting to be afraid

A parting word

A parting challenge

Two Propositions

100 changes

6 change-related issues

100 > 6

Proposition #1

“Ways in which things go right are special cases of the ways in which things go wrong.”

Proposition #1

Successes = failures gone wrong

Study the failures, generalize from that.

Potential data sources: 6 out of 100

Proposition #2

“Ways in which things go wrong are special cases of the ways in which things go right.”

Proposition #2

Failures = successes gone wrong

Study the successes, generalize from that

Potential data sources: 94 out of 100

94/100 ?

OR

6/100 ?

What and WHY Do Things Go RIGHT?

Not just:

why did we fail?

But also:

why did we succeed?

Resilient Response

- Can learn from other fields
- Can train for outages
- Can learn from mistakes
- Can learn from successes as well as failures

http://www.flickr.com/photos/sparktography/75499095/

THE END