That Couldn't Happen To Us... Could It?

James Youngman
Google Site Reliability Engineering

Google Ireland Ltd
2007-03-20
Prepared for UKUUG Spring 2007 Conference, Manchester, UK

Issue 2.10

Disclaimer

DISCLAIMER: Nothing herein is warranted or guaranteed. This product is meant for educational purposes only. Any resemblance to real persons living or dead is purely coincidental. This page is netfake enhanced. Void where prohibited. Some assembly required. List each check separately by bank number. Batteries not included. Contents may settle during shipment. Use only as directed. No other warranty expressed or implied. Do not use while operating a motor vehicle or heavy equipment. Do not look into laser with remaining eye. Postage will be paid by addressee. Subject to CAB approval. This is not an offer to sell securities. Apply only to affected area. No postage necessary if mailed in the United States. Please remain seated until the ride has come to a complete stop. Breaking seal constitutes acceptance of agreement. For off-road use only. As seen on TV. One size fits all. Many suitcases look alike. Contains a substantial amount of non-tobacco ingredients. Colors may fade. We have sent the forms which seem right for you. I did not write the software I am talking about. Slippery when wet. For office use only. Not affiliated with the American Red Cross. Drop in any mailbox. Edited for television. Keep cool; process promptly. Post office will not deliver without postage. List was current at time of printing. Return to sender, no forwarding order on file, unable to forward. Not responsible for direct, indirect, incidental or consequential damages resulting from any defect, error or failure to perform. At participating locations only. Not the Beatles. Penalty for private use. See label for sequence. Substantial penalty for early withdrawal. Do not write below this line. Falling rock. Lost ticket pays maximum rate. Your canceled check is your receipt. Add toner. Place stamp here. Avoid contact with skin. Sanitized for your protection. Be sure each item is properly endorsed. Sign here without admitting guilt. Slightly higher west of the Mississippi. You must be present to win. No passes accepted for this engagement. No purchase necessary. Processed at location stamped in code at top of carton. Shading within a garment may occur. Use only in a well-ventilated are. Keep away from fire or flames. Replace with same type. Approved for veterans. Booths for two or more. Check here if tax deductible. Some equipment shown is optional. Price does not include taxes. No Canadian coins. Not recommended for children. . List at least two alternate dates. First pull up, then pull down. Call toll free number before digging. Driver does not carry cash. Some of the trademarks mentioned in this product appear for identification purposes only. Objects in mirror may be closer than they appear. Record additional transactions on back of previous stub. Unix is a registered trademark of AT&T. Do not fold, spindle or mutilate. No transfers issued until the bus comes to a complete stop. Package sold by weight, not volume. Your mileage may vary. for any utility. Provided as is, with no express or implied warranty, except that provided by the law. If you don't like all this, parents can exercise your discretion. Simulated Picture. Photograph enlarged to show texture. These are performed by professionals. Do not try this at home. Not a toy, keep far away from children. If you are a child, ask your parents to keep you away from this .. and yes .. either parent will do, there no need to get both to do it, so you have no excuse for using this as a toy. Not a spermicide. No user serviceable parts inside. Conforms to FCC part B specifications for Spurious emissions. Fasten seat belts. Its not a good idea, its the law. Recycle and save the world. This mail written entirely with recycled electricity, 100%post-consumer, with vegetable inks. I can't say how many machines we have. Void when printed on any printer. Whatever is not yet degraded is Bio-degradable. 99%cholesterol free. with 70%less fat than Land-o-lakes. You can't believe its not butter .. you don't have to. Non refundable, non transferable. The first washable waterproof mascara. The surgeon general has determined lots of things. This is not a joke .. if it was, you'd be laughing. This message may not be copied in any form of cover other than that originally sent in, and without such a condition being imposed on the subsequent copier. Including all color copiers too ... and remember, I did tell you that one of the guarantees above gets void when you print or fax. Not responsible for clothes left behind. Clothes left in dryer may be removed by next customer. Clothes right in the dryer may not be. Do not put clothes soaked in gasoline in the dryer. Gone for Lunch. No admission without permission. Good for your skin .. tested in a Swiss lab. No added salt or sugar. 0%sodium (acc. to the FDA, 0%sodium if less than so much sodium) .. uses potassium and radium instead. Void where prohibited. Offer expires tomorrow. This program was recorded live and edited for brevity. As seen on TV. Real Psychic - Don't sue, I'd know first. We don't care. Real bagel. real bagels are made fresh by hand. Machine made - untouched by hand. No preservatives. No artificial ingredients. All natural. the synthetic vanilla has been made from naturally occurring coal, methane, and organic compounds. Freshly reconstituted from concentrate. This unit not tagged for individual sale. Are you reading at all? This tag not to be removed under perjury of the law. Murder scene - Do not cross this line. Wheelchair accessible. Politically correct. This perm is guaranteed for life. For the life of the perm. Sale of cigarettes to persons below 18 prohibited, so if you are 18 don't push that special button that is low down and within your reach. We had to put it there to make this unattended vending machine wheelchair accessible, to comply with ADA. Offer not valid with any other offer. No expiry date. Coupon valid until actually presented. No more timeshare presentations. You aren't required to like my talk. Don't ask, don't tell, don't pursue. Mandatory 5 day waiting period - waived for cash. Needles and rubbers perhaps, but Bullets will not be distributed in schools. If verbal consent is not obtained in triplicate, it is date rape. Asking permission constitutes harassment. Between teacher and the taught is unethical and the board distances itself from the statements of the President. Contains some violent scenes. UL certified. Guarantee void if bar code removed. Stop - by opening this you are agreeing to everything. Valid only in continental US - not valid in Alaska, Puerto Rico Hawaii or Canada. Contains cryptographic code. Do not ftp outside the US. All responsibility that of the ftp-er. Do not recharge or swallow. Liability limited to replacement. Void if mixed with other types. This has been written entirely in ASCII. No EBCDIC or animal fat. Coconut Cookies - Two for the price of one Special - No tropical oils. Information on cholesterol is provide for those who are modifying their dietary intake under the advice of their doctor. Perky perky ... why did you read it? Actually none of this was designed to be read. It was meant to be read of course, don't mistake me. Only not designed to be read. Or rather designed to be not read. Directions: Use as desired. All wrongs re-served. All flames to /dev/null. No shirts no shoes no service. Affirmative Action/ Equal Opportunity. Specifications subject to change without notice.

SRE: What We Do

Keep the site up

- Whatever it takes
- Site unavailable? Our problem, whatever the reason

Work at a Large Scale

- Many services
- Lots of data
- Many machines
- But not so many people (machines:admins > 4000:1)

Balance competing demands

- Improve availability and reachability
- Enhance functionality
- Improve efficiency
- Take on new services (post-launch)

How to manage all that stuff?

Have a machine naming convention

- Not enough dwarves
- Or planets
- Or elements
- Star names are difficult to spell

Use a database to store information about machines.

- Hardware configuration
- Software configuration
- Repair history

Automate more, do less.

- Writing scripts is more fun than editing /etc/fstab

Make the computers do the boring stuff

Make The Computer Do It For You

Monitor production systems

- Alert when they fail

Manage the machine database

- Determine physical configuration programmatically
- Update the machine database following upgrades

Detect hardware failures

- Fan stopped, bad memory, disk died, ...

Detect software failures

- Server not running, wrong version, slow response, ...

Apply policy

Heat the office

But My Pager Keeps Going Off

Suppose you have 20,000 machines...

- A number will probably go wrong every day
- Checking that many machines is too time consuming
- Automated fail-over is essential (at or above the level of each possible failure)

Machines break all the time

- Or they just look like they might be broken.
- Figuring out what's wrong is time-consuming, too.
- If we leave them out of service pending diagnosis, we will run out of replacements

The machine database knows what the machine is for

- Use the information to automate problem diagnosis
- Diagnosis? Only half the problem. Automate the repair too!

... and so the pager falls silent.

- More time for foosball

My pager is silent, what do I do?

Wow, we're in Nirvana.

- Manage lots of machines, lots of services
- Monitored automatically
- Automatic failover
- Mostly, fixed automatically

So, what are they paying us for again?

- Capacity Planning
- Some problems the computer can't diagnose or can't fix
- Other things to work on
  - Product features
  - Performance engineering, infrastructure improvement
  - Product and feature launches
  - Guitar Hero
- Pick the task you enjoy least -- and automate it away
  - Now you can work on stuff that's more fun!

Aiming Higher

- It's good, but could it be better?

Going Further

What are the causes of faults?

- Software bugs
- Out of date or incorrect configurations
- Landslides
- Disk failures, broken fans
- Assembly problems
- ...

Now we're detecting hardware problems

- Somebody still has to fix them though
- Some repairs are more urgent than others

A four-disk machine with a broken disk:

- Still 75% working :)
- Repair would take 3 disks out of service
- Repair can probably wait
- There is probably an ideal repair threshold that...
  - Minimises effort spent on repair
  - Maximises the number of in-service disks

Example: Powering disks off

Machines stay in service with broken disks

- Until we're ready to repair them

This causes a number of difficulties

- Extra power usage
- Additional heat output
- Occasional bus resets affect responsiveness and maybe throughput

Powering the disks off will help

- Reduce power usage and heat production
- No more bus resets
- Kernel support for this already exists

Implementing the feature

- Monitoring already exists (supporting repair process)
- Modify the monitor to power broken disks down too

Rolling the change out

Test the change before deploying it

- Obviously we need to do this

Testing and Deployment Plan

- Test in development environment
- Test on a production test system
- Happy? Then...
- Test on a sample production rack
- Test on a volunteer data centre
- Test on some more data centres
- Full roll-out

It works...

Monitoring the tool in operation

- Failed disks get unmounted
- Disks are spun down
- Proven for N weeks in Y data centres
- So, all's well.

Roll it out to the rest of production

- But not all systems have the same hardware
- Manufacturer X disk controllers not fitted in any of the canaries
- Requesting the spin-down of one disk actually spins them all down :(
- That's unexpected
- Kernel panic. Reboot.
- ... on over 50 machines
- ... and the kernel panic causes corrupted local filesystems
- ... which causes data to be under-replicated
- ... GFS pushes out further replicas

Well, at least the problem is contained...

A few machines got rebooted

- But GFS chunk replication ensures no data loss

Investigating the problem

- What had caused the reboots?
- Working on it...

Meanwhile, the corrupted disks are marked as bad

- And spun down, pending repair
- ... causing another reboot ...

Life goes on mostly as normal

- Tools check machine configuration against the machine database
- Unfortunately, one tool checks system configuration against machine DB
- ... and makes the wrong call in this case, updating the machine database!
- These machines appear to have only N-1 disks
- Send them all to repair!

Hang on, where did all the storage go?

Many machines sent to repair

- The good disks in the in-repair machines are no longer available
- Free space in affected GFS cells falls
- Automated repair dispatch suspended when cells are nearly full
- So the repair capacity is taken up with machines which aren't broken...

But at least the fix is simple

- Roll back the spin-down change
- Don't try to spin down disks on Manufacturer X controllers

Er, *fairly* simple

- Fix all the machine DB entries that had been “corrected” to show N-1 disks

Making sure it won't happen again

- Modify the test process for monitoring/repair systems
- Test on x% of machines at a time, across the fleet
- Not just a selected data centre!
- Increase x over time
- Hence detect problems before they do serious damage

Example: Protecting data with checksums

Large data volumes

- Pushed over the network
- Replicated via GFS
- Stored on disk

Component specs tell us that bit error rates are nonzero

- At these data volumes, *expect* some corruption
- So, use checksums to detect this situation
- Checksums *don't protect* data, they tell you it's *already* broken
- Ensure that there is a way of recovering from the problem (i.e. have several copies of the data)

Checksums

- IP and TCP already have checksums
- Ethernet does too
- Store checksums on-disk for GFS data
  - GFS is high-bandwidth, use an algorithm which protects data at low CPU overhead
- Various options for data in memory

Defence In Depth

The implementation is important

- IP checksum protects only the header
- TCP checksums are only 16 bits
- If Ethernet frames are modified as they pass through a device, the CRC is recomputed
  - What about the possibility of data corruption while in the network device?

Use checksums at the application level

- Provides end-to-end protection
- No longer relying on the network to protect in-transit data
- Pick a checksum algorithm which is fast enough that high-throughput applications don't grind the CPU if they're just doing I/O

What to do for checksum failure?

- Could just discard the data – viable only for some situations
- If a GFS client receives data from a chunk server with an invalid checksum
  - The problem could be in the client, the server, or the network
  - Try again, try another chunkserver
  - Report the problem to a central point (aggregate to diagnose systematic problems)

Cue Ominous Music...

Processes start dying in one data centre

- Application is reporting fatal data verification errors
- GFS data checksum mismatch
- Checking the GFS files manually shows the application is right
- Huge numbers of GFS checksum errors reported in this cell in a two-hour period
- Correlate the data
- The chunk servers affected are pretty much all in the same rack
- Take the whole rack out of operation

Analyse the nature of the corruption

- Compare different replicas of the data
- Bit flips
- Testing with scp shows that they are caused by a broken switch
- Dense enough to fool the TCP checksum
- In fact, they are double-bit flips!
- The checksum algorithm in use (Adler32) turns out to be ineffective in protecting us from these (see also RFC 3309)
- Despite the 2x performance difference, better to use CRC than Adler32

Summary

Both problems caused by undesirable interactions

- Both of our example problems feature interactions between hardware and software design
- Often, the problems that bite most painfully are the ones involving complex interactions

Why?

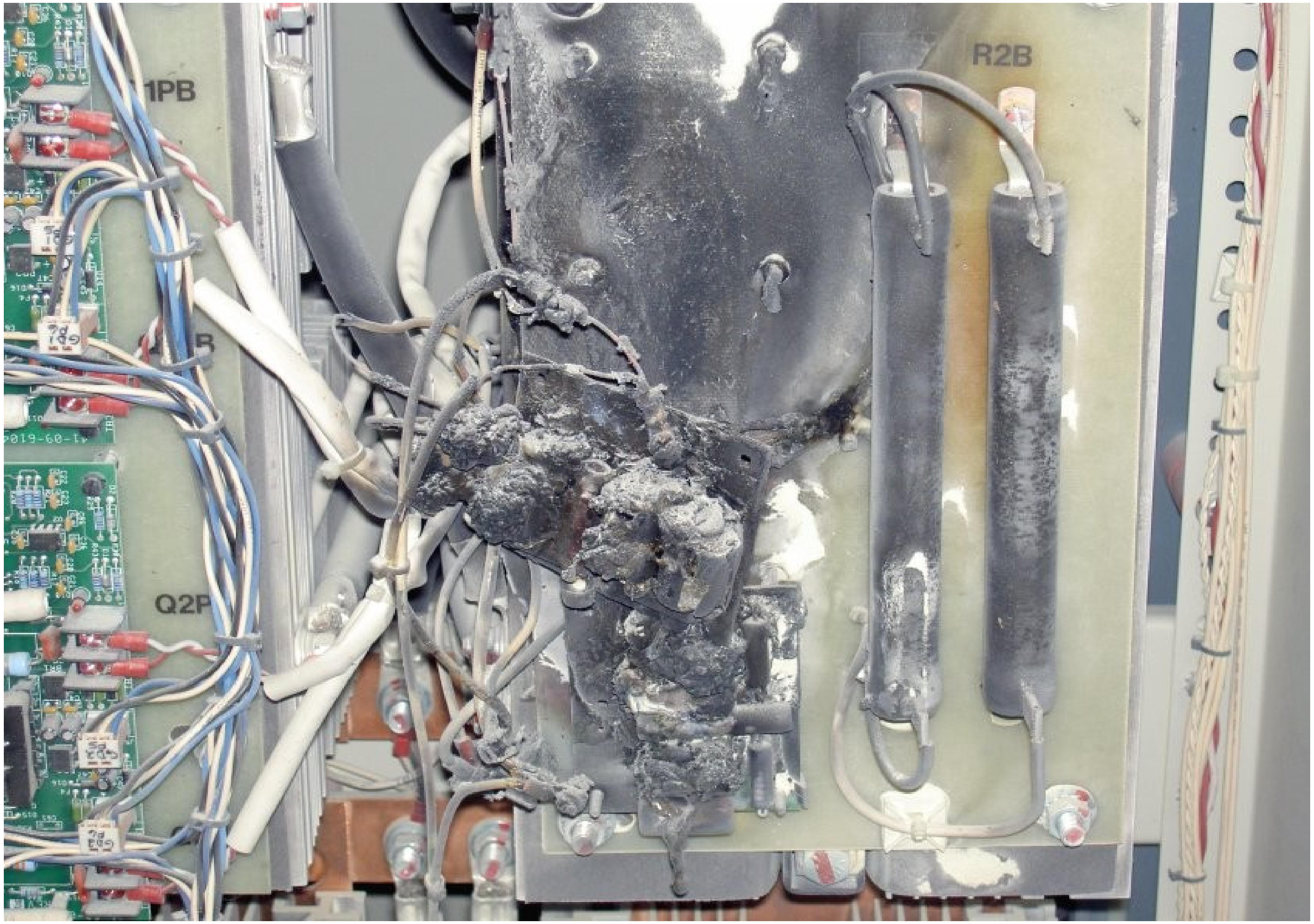
- Because problems which don't involve complex multi-component interactions are much easier to find during design reviews, during testing, and so on
- Because these scenarios are hard to reproduce
  - For example, how often does pre-production testing happen on known-broken hardware?

Lessons Learned

- Devise testing strategies that cut across every type of variability
- Automation saves a lot of manual effort, but it's a bigger hammer
  - It hurts more when you whack your thumb with it
- Redundancy prevented data loss and end-user impact in each case
  - Redundancy and transparent fail-over at (or above) every level is essential

And lastly...

Ouch



Ouch, again