CS6501-015
CLOUD SYSTEM RELIABILITY

Course Introduction

Prof. Chang Lou, UVA CS, Fall 2023
AGENDA

— What this course is about
  — why study on cloud system reliability?

— What you can expect from this course
  — and what we expect from you

— But first, tell us more about yourself!
  — who you are, and why you took this course
WHY YOU TOOK THIS COURSE

– Likely answers

– No exam
– Interested in cloud computing
– Get experience for related jobs in industry
– Try a small research project
– Write more reliable codes
WHY CLOUD SYSTEM RELIABILITY MATTERS
WHY CLOUD SYSTEM RELIABILITY MATTERS

— Modern world depends on cloud systems
WHY CLOUD SYSTEM RELIABILITY MATTERS

– In 2023, 94% of enterprises use cloud services

WHAT IS RELIABILITY

— What are some common qualities we measure on systems?
WHAT IS RELIABILITY

— Reliability is not
  — Performance: make systems faster
  — Usability: make systems more user-friendly
  — Security: make systems safer against intrusions
  — Cost-effectiveness: make systems more affordable

— Reliability is
  — the system's ability to consistently perform its intended function without failure over a given period.
CLOUD FAILURES

– Cloud failures are prevalent
— Cloud failures can be really annoying

*Sgt. Brink @LASDBrink*

#Facebook is not a Law Enforcement issue, please don't call us about it being down, we don't know when FB will be back up!

Reddit when youtube's been down for 5 min

*Joe Brown @joemfbrown*

I'm sitting here in the dark in my toddler's room because the light is controlled by @Google Home. Rethinking... a lot right now.

*Everybody right now.*

#AWS #awscloud #awsoutage #awsdown #S3 #AWSs3 #Amazon

*OLD MAN YELLS AT CLOUD*
CLOUD FAILURES

—.. or much worse, huge economic loss and service unavailability

Microsoft’s MFA is so strong, it locked out users for 8 hours

3 difficult days for Rackspace Cloud Load Balancers

Posted by iwcr

Amazon ‘missed out on $34m in sales during internet outage’

The e-commerce giant generates $9,615 in sales per second – but not when it’s website is down

Ben Chapman • Tuesday 08 June 2021 16:54 • 1 Comments

Millions online hit by Microsoft 365 outages

After almost 24 hours of technical difficulties, Facebook is back

Facebook blamed the issue on a “server configuration change.”

911 emergency services go down across the US after CenturyLink outage

Zack Whittaker @zackwhittaker / 4 months ago
CLOUD FAILURES

– Cloud systems fail due to different root causes

Google Fiber Shot Down By 'Bored' Hunters

'Bored' Hunters Shoot Down Google Fiber

By Bianca Bosker
Nov 22, 2010, 05:12 AM EST | Updated May 25, 2011, 05:50 PM EDT

Google reinforces undersea cables after shark bites

Sharks have been biting down on fibre optic cables under the Pacific, possibly confused by electrical signals that resemble fish
IN THIS COURSE

Challenges
- software bugs
- misconfigurations
- hardware faults
- network issues
- human mistakes
- scale
- ...

Techniques
- program analysis
- formal methods
- diagnosis
- ...

GOAL OF THIS COURSE

— Expose students to system reliability research
— Get students familiar with key concepts and techniques
— Provide a research opportunity for students to experience
IN THIS COURSE

— Review literatures about cloud reliability
  — classic work + state-of-art
  — from system top conferences: SOSP/OSDI, NSDI, ASPLOS..

— Learn how to present a research work
  — ..and defend it like you were the author!

— Explore a research topic you feel excited ★
  — a try-out for you if you are interested in Ph.D.
Course Info
COURSE INFO

- **Time/Location**
  - TuTh 12:30pm - 1:45pm, Rice Hall 011

- **Office Hours**
  - TuTh 4:00pm - 5:00pm, Rice Hall 304

- **Discussion Forum**
  - UVA Canvas (where you submit reviews)

- **Questions**
  - Send emails to cs6501-015staff@virginia.edu or stop by at my office (open-door policy)
LECTURER INFO

— Prof. Chang Lou
  — Ph.D. from Johns Hopkins
  — Joined UVA CS as a faculty member in Fall 2023

— Research Areas
  — Distributed Systems, Operating Systems, Software Reliability

— Research Goal
  — Enhance cloud systems to be more resilient against arising reliability challenges
$ whoami
   - Wonkyo Choe       # often go by “Won”
   - third year Ph.D. student at UVA CS
   - working w/ Prof. Lin

$ echo $OFFICE
   - Rice 430

$ head research_interests
   - OS – The Linux Kernel
   - System for ML

$ cat ta.email
   - wonkyochoe@virginia.edu

You may just want to see a cute cat instead of me
GRADING

— Reviews: 15%
— Class Participation: 15%
— Presentation: 20%
— Project: 50%
REVIEWS

– Must read both papers before each class

– Prepare an one-page review for one reading assignment
  – choose Reading A or Reading B

– Submit to Canvas before 10:00 am on class day
  – maximum three reviews are allowed to miss without penalties

Violation of UVA Academic integrity: Directly copying from paper contents, peers or online resources will be considered as a violation of academic integrity and will lead to consequences.
LATE POLICY

— Everyone has 96 hour late tokens (for reviews, report ...)
— To use, just send an email to staff email list
   — late submissions using late tokens receive no penalty
— What if I run out of tokens?
   — 1 day late, 15% deduction
   — 2 days late, 30% deduction
   — 3 days late, 60% deduction
   — after 4 days, no credit.
CHATGPT POLICY

— How to best use AI writing assistant
  — brainstorm initial ideas
  — check grammar errors

— You should not use AI to
  — directly generate reviews for you

It is a violation of academic integrity as well.
HOW TO READ A PAPER (3-PASS APPROACH)
by S. Keshav, University of Waterloo

— The first pass: general idea (5-10 mins)
  — title, abstract, and introduction
  — headings
  — conclusion

— The second pass: content (< 1 hour)
  — figures, diagrams and other illustrations in the paper
  — references

— The third pass: details (may take hours)
  — "re-implement" the paper

https://web.stanford.edu/class/ee384m/Handouts/HowtoReadPaper.pdf
HOW TO READ AN ENGINEERING RESEARCH PAPER
by William G. Griswold, CSE, UC San Diego

— What are the motivations for this work?
— What is the proposed solution?
— What is the work's evaluation of the proposed solution?
— What is your analysis of the identified problem, idea and evaluation?
— What are the contributions?
— What are future directions for this research?
— What questions are you left with?
— What is your take-away message from this paper?
CLASS PARTICIPATION

— Attend classes
  — discuss reviewed papers with peers
  — there will be a few random (and simple!) quizzes 😊

— Exceptions: illness and other absences
  — If you are ill, you should stay away from classes
  — or in need to attend career development events - interviews, conferences, ..
  — three absences are allowed with no questions
PRESENTATION

- Everyone needs to present twice through the semester
  - Register at Google Sheet TODAY! (Link in Canvas Announcement)
  - Presenters in First Eight Slots get bonus credits

- For each presentation
  - Main body: 25-30 min
  - Q&A: 5 min

- Lead the discussion and defend the work
  - (it is called thesis "defense" for a reason)
PROJECT
STEPS OF A RESEARCH PROJECT

1. Problem Identification
2. Proposing Hypothesis
3. Review of Related Literature
4. Preparation of Design
5. Experimentation
6. Results and Discussion
FIND YOUR TOPIC

— Approach 1. Address the limitation of an existing paper
  — "in this work we made the assumption that ..."

— Approach 2. Revisit classic problems in new scenarios
  — e.g., reliability problems in serverless/micro-services

— Approach 3. Find inspirations from external sources
  — what is the problem people complain about but no good solution yet?
INSCRIPTIONS: HACKERNEWS

https://hn.algolia.com/?q=reliability
One of my favorite pastimes is lunchtime with software engineers where I like to pose the following question: on average, how many software defects exist per 1,000 lines of delivered code? I've gotten answers across the board, and research backs up similar results - answers vary widely. But no one says “none.” Results are subject to many factors, from programming language to developer skill, but some seem to suggest it is in the range of 10-20 per 1,000 lines.

For comparison, I just finished a small piece of C-code totaling 780 lines. Once all the obvious errors were dealt with and it compiled, the memory checker found four more serious issues. That's roughly five in 1,000 lines – but those are just memory issues like uninitialized memory reads and array overruns by one, etc. It does not indicate completeness, correctness, or even testing for infrequent code paths not included in the run. 10-20 per 1,000 seems right on the money to me.

The problem is cryptography may be mathematical algorithms, but the math is implemented in code. Bugs in cryptographic code happen. Recall the Java 15+ certificate validation bug? The bug was failing to check that the integers used in the algorithm were of sufficient size. In fact, integers of 0 qualified just fine, and so a certificate of all zeroes would be accepted as valid for whatever identity you tried to assume. Big bug, big consequences.
# INSPIRATIONS: POST MOMENTUM

---

## Google Cloud Service Health

This page provides status information on the services that are part of Google Cloud. Check back here to view the current status of the services listed below. If you are experiencing an issue not listed here, please contact Support. Learn more about what's posted on the dashboard in this FAQ. For additional information on these services, please visit [https://status.cloud.google.com](https://status.cloud.google.com).

### Available | Service information | Service disruption | Service outage


Multiple Google Cloud Platform services impacted globally with operational latency

Incident began at 2023-06-12 17:15 and ended at 2023-06-12 21:48 (all times are UTC/Pacific)

Previously affected location(s):

- Tokyo (asia-east1), Hong Kong (asia-east2), Tokyo (asia-northeast1), Osaka (asia-northeast2), Sendai (asia-northeast2), Seoul (asia-south1), Busan (asia-south2), Singapore (asia-southeast1), Jakarta (asia-southeast2), Sydney (australia-southeast1), Wellington (australia-southeast1), Sydney (australia-southeast2), Warsaw (europe-central2), Helsinki (europe-north1), Madrid (spain-southwest1), Belgium (europe-west1), Austria (europe-west2), London (europe-west1), Frankfurt (europe-west6), Amsterdam (europe-west3), Milan (europe-west2), Paris (europe-west3), Delft (eu-central2), Malmo (europe-southwest1), Stockholm (europe-northwest1), Helsinki (europe-northwest1), Lisbon (europe-west1), Cape Town (southafrica-west1), Iowa (us-central1), South Carolina (us-east1), Northern Virginia (us-east4), Columbus (us-east2), Dallas (us-south1), Oregon (us-west1), Los Angeles (us-west2), Salt Lake City (us-west3), Las Vegas (us-west4)

<table>
<thead>
<tr>
<th>DATE</th>
<th>TIME</th>
<th>DESCRIPTION</th>
</tr>
</thead>
</table>

We thank you for your patience while we worked on resolving the issue.

Summary: Multiple Google Cloud Platform services impacted globally with operational latency.

- Description: Mitigation work is still underway by our engineering team.
- We will provide more information by Monday, 2023-06-12 22:45 UTC/Pacific.

### Diagnostic:

- Users may observe failures or delayed operations for affected services.
- Google Compute Engine Impact: Diagnostic users may experience elevated latency for API calls to shared, regional and zonal resources.
- Cloud Memorystore Impact: Diagnostic users may be unable to create new node instances, and likely not be able to delete instances.
- Google Cloud Composer Impact: Diagnostic users may observe failures while running create operations using Cloud Composer.
- Google Cloud Dataflow Impact: Diagnostic impacted customers may see increased latency during WorkerPool start up and for Dataflow jobs.

---

https://status.cloud.google.com/incidents/VuCtCwkRXueAyusvrXfG
What if I cannot think out of anything?

It's fine, we have a list of prepared topics

but we suggest coming up with your own idea, something you feel passionate about

a fun and novel idea with okay execution is better than a boring idea executed perfectly
Exercise: bug detection tool for ChatGPT generated codes in distributed systems
SOURCES: JIRA

https://issues.apache.org/jira/browse/ZOOKEEPER-3832

ZKHostnameVerifier rejects valid certificates with subjectAltNames

Details
Type: Bug
Status: OPEN
Priority: Major
Resolution: Fixed
Affects Version/s: 3.6.1, 3.6.8
Fix Version/s: 3.5.9, 3.7.0, 3.6.2
Component/s: server
Labels: None

Description
This is the same issue as reported in
https://issues.apache.org/jira/browse/HTTPCLIENT-1906
For performance reasons we use a copy-and-pasted version of the HostnameVerifier, as a consequence we don’t pick up these fixes automatically.
### SOURCES: BUGZILLA

**Bug List**

- Component: ext4
- Product: File System

<table>
<thead>
<tr>
<th>ID</th>
<th>Product</th>
<th>Component</th>
<th>Resolution</th>
<th>Summary</th>
<th>Changed</th>
</tr>
</thead>
<tbody>
<tr>
<td>4556</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>odd: kernel BUG at SchBuffer.c:39501 Init by extent 209</td>
<td>2013-04-20</td>
</tr>
<tr>
<td>69701</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>odd: [kthreadd] fs_ext4</td>
<td>2013-11-26</td>
</tr>
<tr>
<td>9955</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>filesystem should reserve modes for stat as they do disk space</td>
<td>2015-05-16</td>
</tr>
<tr>
<td>69712</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>out of fileset detected in function ext4_file_open</td>
<td>2016-09-23</td>
</tr>
<tr>
<td>72125</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>ext3: bad_extent:14704 competing in bitmap: 248416 of fs</td>
<td>2016-03-20</td>
</tr>
<tr>
<td>72201</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>EXT4 fs mount (device 8): ext3: bad_extent:14704 at bitmap: 248416 of fs</td>
<td>2016-03-20</td>
</tr>
<tr>
<td>72392</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>src: memory failure on new indents read &amp; unindent array</td>
<td>2016-03-20</td>
</tr>
<tr>
<td>76641</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>ext3: da_adapters etc: 30 after mount (in directory)</td>
<td>2016-03-20</td>
</tr>
<tr>
<td>78721</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>WARNING: CPU: 2 PID: 1512 fs/ext4:ext3:216 when mounting an external disk</td>
<td>2016-03-20</td>
</tr>
<tr>
<td>89521</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>ext3: mount I don't work as ext4?</td>
<td>2016-03-20</td>
</tr>
<tr>
<td>80551</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>write performance of ext4 improve (write as volume file)</td>
<td>2016-03-20</td>
</tr>
<tr>
<td>80540</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>Chroot root entries are possible to create in a file</td>
<td>2016-03-20</td>
</tr>
<tr>
<td>80559</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>RPM: /usr/lib64/10011, blocked for more than 100 seconds.</td>
<td>2016-03-20</td>
</tr>
<tr>
<td>87612</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>BAD_interconnect cause:0x21 at blockoffsets when used as a mounted ext2, fs, fs1</td>
<td>2016-03-20</td>
</tr>
<tr>
<td>88321</td>
<td>File Sys</td>
<td>fs_ext4</td>
<td>NEW</td>
<td>WARNING: CPU: 1 PID: 6784 at bitmap: 1:018 amount, check fs/ext4/2727</td>
<td>2016-03-20</td>
</tr>
</tbody>
</table>

https://bugzilla.kernel.org/buglist.cgi?component=ext4
SOURCES: GITHUB ISSUES

https://github.com/rust-lang/rust/issues
FIND LITERATURES

– From top system conferences
  – SOSP/OSDI/Eurosys/ATC (system)
  – HotOS (workshop)
  – NSDI/SIGCOMM (network)
  – ASPLOS (programming language and arch)
  – FAST (storage)
  – SOCC (cloud)
PROJECT

— Week 1: Form a team (2-3 students)
  — brainstorm on your project topic (related to cloud system reliability)
— Week 2-3: Schedule a meeting with me to discuss
— Week 4-5: Write and submit proposal (1-2 pages)
— Week 6-14: Do research
  — Week 11: Checkpoint report due (~3 pages)
— Week 15-16: Presentation
— Week 17: Final report (including codes)
IMPORTANT TIMELINE

8/29  Team registration due
9/21  Proposal due
11/2  Checkpoint report due
11/30 Presentation 1
12/5  Presentation 2
12/12 Final report due
— Reliability
  — the probability that a system operates without failure in a given period of time.
  — how to compute probability: Mean Time Between Failures (MTBF)

\[
Reliability = 1 - \frac{1}{MTBF} = 1 - \frac{\text{Num of Breakdowns}}{E[uptime]}
\]

— Availability
  — the percentage of time that the system operates satisfactorily.

\[
\text{Availability} = \frac{E[uptime]}{E[uptime] + E[downtime]}
\]
RELIABILITY VS AVAILABILITY

— Reliability and availability are related concepts
— but not same! Imagine there are two bad laundry machines:

Stop working every 6 months, takes 1 week to repair

Stop working every 3 weeks, takes 1 hour to repair

— Modern distributed systems: unreliable but high available
THE BATHTUB CURVE
RELIABILITY VS SECURITY

— Both are essential aspects of system design
  — reliability and security research have a lot of similarities
  — many bugs can both hurt system reliability and security

— Key difference: adversary
  — reliability: bad designs and mistakes are their own worst enemies
  — security: threats with bad intentions (malware, hackers, etc.)
FAULT, ERROR, FAILURE

fault
root cause: a bug, misconfiguration, hardware issue, etc.

error
fault triggered, sth goes wrong (e.g., I/O error), could be handled

failure
visible impact to users
Q & A

– ToDo 1) Team up
  – find your teammates and discuss ideas
  – team leader emails me name list and schedules meeting (ddl: 8/29)

– ToDo 2) login Canvas
  – sign up the presentation schedule for two classes
  – submit review for Thursday class

chlou@virginia.edu
ACKNOWLEDGEMENT

PROF. RYAN HUANG (UMICH) AND PROF. TIANYIN XU (UIUC) FOR SHARING THEIR COURSE MATERIALS AND TEACHING INSIGHTS