Chaos Engineering Building confidence in your application and team through failure experimentation

Christopher Meiklejohn

October 29, 2020



Carnegie Mellon University School of Computer Science



Homework 4B was due Nov 3rd, **now the 4th**.

Homework 4C was due Nov 5th, **now the 6th**.



Carnegie Mellon University School of Computer Science

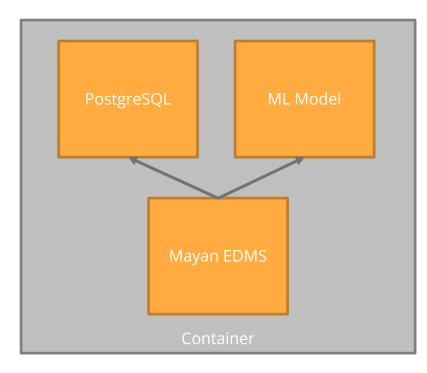
Learning Goals

Identify the need for chaos and resilience engineering

Understand the principles of chaos engineering



Exercise: Monolithic Application



What kind of failures can happen here?

How likely is that error to happen?

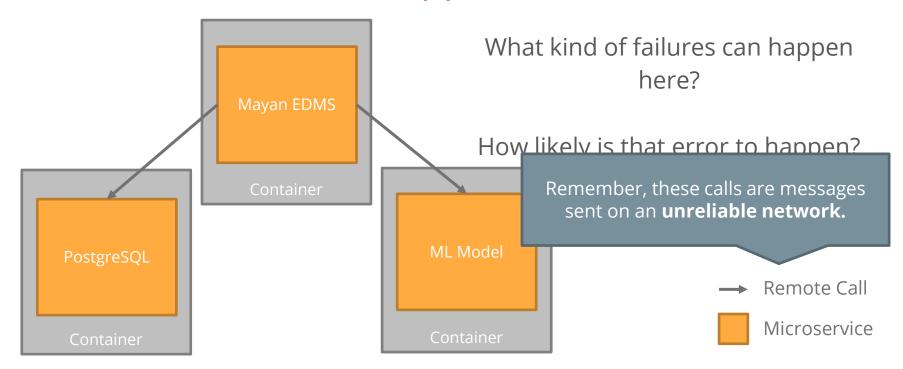
How do I fix it?







Exercise: Microservice Application





Failures in Microservice Architectures

- 1. Network may **be partitioned**
- 2. Server instance **may be down**
- 3. Communication between services may **b**

All of these issues can be indistinguishable from one another!

Making the calls across the network to multiple machines makes the probability that the system is operating under failure **much higher.**

- 4. Server could be overloaded and respon
- 5. Server **could run out of** memory or CPU

These are the problems of **latency** and **partial failure.**



Where Do We Start?

How do we even **begin to test these scenarios?**

Is there any **software** that can be used to test these types of failures?

Let's look at a **few ways** companies do this.



Carnegie Mellon University School of Computer Science

Game Days

Purposely **injecting failures** into critical systems in order to:

- Identify **flaws** and "latent defects"
- Identify subtle dependencies (which may or may not lead to a flaw/defect)
- Prepare a **response** for a disastrous event

Comes from "resilience engineering" typical in high-risk industries

Practiced by Amazon, Google, Microsoft, Etsy, Facebook, Flickr, etc.







mine with desse fielders, Kope Scholter, Ants Allipper, and Son (dessecut) swally die lufdig: diopping source and searching is very source at a Add creater bandling

transaction for our of the longert online strait operations in the coastry. Some systems have folial, and on new tensor why, force lower our of the shares while lowes of organizers well assumd the dash for three days trying in storage. The and speciel to that it's not a well also days, offerein it could have lowe, instead, it's an everyte derived

In tasket a conject low to adapt to the alectedne vision (adapt. Thus proce, denote happen, hadne to out, difficulty, and some sprinterly lack of all therain adapted and advance registrem. *Beil to Market Wer* and yoursel (dates, the lot properties) (if all a feet a sample fit. However, persistence are shortly or how the composition of the logical "discover dependence" to follow the where the however better add to adapt to it where it comes. Environment optical question is to follow, the out is an eventual wave and and an event of the logical date of the logical date of the logical date of the logical with the however better add to a date in it where it comes. Environment optical question is the logical date of the induction of a southern and the logical dates on advances. The logical dates and the logical dates and the logical dates are induced and the logical dates and the logical dates and the logical dates and the logical dates and the logical dates.

etc. As the only XMM, down result (Cancelley 4 prepare dispect) interface collisist by particular inputs more futures into colors a conset serving and or 6 down them and which appendix to the transfer of colors works for a compared by streme, of them and particle from one of partners for a toporter to a down one constraints are prepared as prepared for the constraints of the other a few same, but were of the acceleration.

An encourse testing the second second

prior formers. Graph operation of Gamedian locate by Krype Krishware, who has been with the program advant free by these if started at some app. See date works on other industriction practice, must of which are forward on the sometime of some of which date.

or principal or seen and new state. Mainstrate this document in a charget unliques. These Lancourelly, where is a new enhabitity expresses Millionees in system admenticular costs, for shareds and an Ard Later and his section Sine heads, second codelly. These Makagement for System Addicationamen and The Practice of System and Network administrations.

TOM LINOXCELLI Sone, you've periodsly been involved in more GameDay combine than ambody. What's the course important issues you've taken away from that?

Carnegie Mellon University School of Computer Science

- Goals:
- **Preemptively trigger** the failure, observe, and fix the error
- Script testing of **previous failures** and ensure system remains resilient

Failure is inevitable (fraction of percent; at Google scale, ~multiple times)

Build the necessary relationships between teams **before** disaster strikes

Game Days

Our applications are built on and with **"unreliable"** components





Example: Amazon GameDay

Full data center destruction (Amazon EC2 region)

Not all failures can be actually performed and must be **simulated!**

ta center will be taken offline er will be taken offline

- Jonny and the house (months) that a GameDay will be happening
- Real failures in the production environment

Discovered **latent defect** where the monitoring infrastructure responsible for detecting errors and paging employees **was located in the zone of the failure!**





the second with down fielded, Krige Withden, Alth Allipate, and fee Universit 1 way weatly the building dropping second and seconding is very strong at a data credit face. Building

transactions for our of the larged author studie positions in the country. Some systems have failed, and an new interest why, faces larged an off the sharts while issues of organizers and annual the dask for three large trying to strategy. The galaxies and alkalar-shough it could large items. Indexd, it is an eventor despeed The gala speed set of that its out a stud disadar-shough it could large items.

In both a compare loss or adapte of the short-failer series radius. Thing print, almost happen, but in the real difficult mean sequences that the adapte of the print of the print of the print of the print bases and print print print of the print print print of the the second print of the print of the print bases and print pri

edit. As the only 2006, diseases constrained and a pagester damped in several resilience by particular ingesting many datases that calculated assess wave regarded to disease of an and allot dependences. Instands, of calculate calculated and an annual calculated and applies that can use of particular a stageness to a disease over the disease annual calculated and the first second and several the tables of the same tables are structured assess to be all other annual structures and the same structure and the same tables of the same tables and the same structure tables of the same tables of the same tables of the same tables of the same tables and the same structure tables and the same structure tables of the same tables and the same tables of the same tables of the same tables and the same structure tables of the same tables and the same tables of the same tables of the same tables and the same tables of the same tables are tables and the same tables and the same tables and the same tables and the same tables are tables and the same tables and the same tables and the same tables and the same tables and tables are tables are tables are tables are tables and tables are tables and tables are tables and tables are tables and tables are tabl

pain kertensi. Gongle zuprinteri di "Laned-bay ternet-by Kolpa Kerlahana, arke her bert wilk the program obsert (her by their et started als pain ago. Ste also works on other infrastructure projects, most of which are found at

(v) presenting in previous previous accurate a complexity of previous and previous and previous distribution of the complexity of previous and the section and rest of the complexity of previous and the transmission of previous distribution and the Practice of Section and Network Internationalism.

TOR LINONCILLE loss, you've periods' loss inschool in more Consider exaction data anglische. What's die coost important losser motive takes anny from that?



inn with Janua Robbins, Kripa Krishkan, John Ellipae, and Ton Unancelli

samactions (in our of the larged orders what operations in the county, Some oppose have field, and on we know why, fixed levels are of the sharts while larges of organism west around the disk for three days

The grad network is blued to real a real abusitier - doings of a shall have these brokeds. It's over entrols integrade to task is compared in the shadper to the morthold's control shallow. Thengs these, is during targets, folder or real. Although no new -perturbative hast of all and these all abust integrates, and targets and targets and the before they can't perturbations, this has to perturbative and grade parts of an experiment of a follow, that Although experiments or attribute or applicative and the larget of herearchip lengthermions for follows, that

erfl. 3 Broach XMD, Innanon reactif Ganel Qu 2 program disputci in increase realistics of programs increases and plateau networks and a contrast series regarded to discuss times and adult approximations increases of catalogue controls in our company controls of them at papels in the conset of approxy a segment to a Assertance react. Plateau acceptance of the Ganel Star control has show a few series, then amount programs on the show approxy and a segment of the Ganel Star control has been after series to a segment to a Assertance react. Plateau acceptance of the Ganel Star constant, Plate Basertance annuality in the series of the set of the set of the set of the Ganel Star constant. The Basertance assertance and the set of the set

Gough's quarteristics of Garad-Day Is man by Krippe Relationse, who has been with the program advant from by these at started six soon ago. She also works on other industriction practics, word of which are focused on

We presenting the discussion is a Graph colleague. These Ensemble (R) which is not endublishy expresses Multilatures in content-admentication coulds, for databation of all Add Lado and has seenting from bands, seent redshifts Tame Mastagement for Spream Administrations and The Tractice of System and Network

TOM EMONCTED Sone, you've preliably been involved in more Generilay combine than anybody. What's the encod important lesser wor've taken away from that?

Yes, and the exercise should be designed to make people feel a little uncomfortable. The truth is that **things often break in ways that people cannot possibly imagine.**



John Allspaw Former CTO Etsy

I've got a crazy story...



Kelly Sommers Cassandra MVP

ISC institute for SOFTWARE RESEARCH Carnegie Mellon University School of Computer Science

Google GameDays

Experiments take roughly 24 – 96 hours:



For one smartly the building disopping seasons and something is view strong at a 2020 credit bandling transactions for one of the longest orders with expensions in the contexty. Some rythme have fields, and an area haven why. These longest area of the shorts of help latents of argument work strongs the data for three data them is reverse.

viet. As the only XXMs, it makes resulted can be participated in increase redining by particular spectra quark plants into choice a consex serving and or a discuss of time, and a data dynamication in transition of canonical events of the comparison of the consect participation of the one of participation is supported to a discussion or each of the discussion of the Linear Data should be show a few same that more comparison wave of the data supports and in addition of the linear Data show a few same than the support of the discussion or each of the data structure of the Linear Data structure on section. The discussion exercision of the discussion of the data subports and addition of the structure section. The discussion exercision.

som of the opticities. Description is sub-the balance that the net high of classes is due to exact the state of the optical o

prior internets. Complex appression of Cannel Jay Invan by Krippe Reichenses, who that here with the program about from by these it started als some age. She also works on other infrastructure projects, must of which are focused an

ny proceeding the activation of a data of the second state of the

ON LINOXCELLE loss, you've preliably loss: involved in new SemiDay exercise due anglody. Out's the root important losser you've taket away from that?

- 00 24h: the initial response, the appearance of the 'big' problems.
- 24 48h: team to team testing and response, bi-directional testing
- 72 96h: exhaustion; part of the test to identify the human response

...in a real emergency, you might not have the option of handing off work at the end of your shift.



Kripa Krishnan Director, Cloud Ops & Site Reliability Engineering Google



Carnegie Mellon University School of Computer Science

AWS Outage: April 21, 2011

EC2 **outage** in us-east-1 (Northern Virginia)

Outage **affects**:

- Foursquare
- Quora
- Reddit

Outage results in **performance problems** and in some cases **data loss**



AWS Outage: April 21, 2011

At 12:47 AM PDT on April 21st, a **network change** was performed as part of our **normal AWS scaling** activities in a single Availability Zone in the US East Region. The configuration change was to upgrade the capacity of the primary network. During the change, one of the standard steps is to shift traffic off of one of the redundant routers in the primary EBS network to allow the upgrade to happen. The traffic shift was executed incorrectly and rather than routing the traffic to the other router on the primary network, the traffic was routed onto the lower capacity **redundant EBS network.** For a portion of the EBS cluster in the affected Availability Zone, this meant that they did not have a functioning primary or secondary network because traffic was purposely shifted away from the primary network and **the secondary network couldn't handle** the traffic level it was receiving. As a result, many EBS nodes in the affected Availability Zone were completely isolated from other EBS nodes in its cluster. Unlike a normal network interruption, this change disconnected both the primary and secondary network simultaneously, leaving the affected nodes completely isolated from one another.



AWS Outage: April 21, 2011

When this network connectivity issue occurred, a large number of EBS nodes in a single EBS cluster lost connection to their replicas. When the incorrect traffic shift was rolled back and network connectivity was restored, these nodes rapidly began searching the EBS cluster for available server space where they could re-mirror data. Once again, in a normally functioning cluster, this occurs in milliseconds. In this case, because the issue affected such a large number of volumes concurrently, the free capacity of the EBS cluster was quickly exhausted, leaving many of the nodes "stuck" in a loop, continuously searching the cluster for free space. This quickly led to a "remirroring storm," where a large number of volumes were effectively "stuck" while the nodes searched the cluster for the storage space it needed for its new replica. At this point, about 13% of the volumes in the affected Availability Zone were in this "stuck" state.

Primary Outage

At 1247 AM PDT on April 21st, a network change was performed as part of our normal AWS scaling activities in a single Availability Zone in the US East Region. The configuration change was to upgrade the capacity of the primary network. During the change, one of the standard steps is to shift traffic off of one of the redundant routers in the primary EBS network to allow the upgrade to happen. The traffic shift was executed incorrectly and rather than routing the traffic to the other route on the primary network, the traffic was routed onto the

lower capacity redundant EBS network. For a portion because traffic was purposely shifted away from the p affected Availability Zone were completely isolated fr network simultaneously, leaving the affected nodes o

When this network connectivity issue occurred, a larg network connectivity was restored, these nodes rapid cluster, this occurs in milliseconds. In this case, becaus many of the nodes "stuck" in a loop, continuously sea while the nodes searched the cluster for the storage s

After the initial sequence of events described above, t entered the re-mirroring storm and exhausted its avai volume AP in particular) was configured with a long t plane has a regional pool of available threads it can u had no ability to service API requests and began to fa disabled all new Create Volume requests in the affect

Two factors caused the situation in this EBS cluster to when they could not find space, but instead, continue to fail when they were concurrently closing a large nu however, during this re-mirroring storm, the volume c bug, resulting in more volumes left needing to re-mir

By 5:30 AM PDT, error rates and latencies again incree EC2 instance, the EBS nodes with the volume data, an replica and recognized by the EC2 instance as the play of the race condition described above, the volume of calls increased as the system retried and new request team began disabling all communication between the affected Availability Zone (we will discuss recovery of

A large majority of the volumes in the degrade BBS team developed a way to prevent BBS servers in the o other essential communication between nodes in the becoming "stuck". Before this change was deployed, t becoming "stuck". However, volumes were also slowly that when this change was deployed, the total "stuck"

Customers also experienced elevated error rates until zone. This occurred for approximately 11 hours, from



Recovering EBS in the Affected Availability Zone

to find enough space to create new replicas.

were able to create EBS-backed EC2 instances but were experiencing significantly-elevated error rates and latencies. New EBS-backed EC2 launches were being affected by a specific API in the EBS control plane that is only needed for attaching new instances to volumes. Initially, our alarming was not fine-grained enough for this EBS control plane API and the launch errors were overshadowed by the general error from the degraded EBS cluster. At 11:30 AM PDT, a change to the EBS control plane fixed this issue and latencies and error rates for new EBS-backed EC2 instances declined rapidly and returned to near-normal at Noon PDT.

RDS multi-AZ deployments provide redundancy by synchronously replicating data between two database replicas in different Availability Zones. In the event of a failure on the primary replica, RDS is designed to automatically detect the disruption and fail over to the secondary replica. Of multi-AZ database instances in the US East Region, 2.5% did not automatically failover after experiencing "stuck" I/D. The primary cause was that the rapid succession of network interruption (which partitioned the primary from the secondary) and "stuck" I/D on the primary replica triggered a previously un-encountered bug. This bug left the primary replica in an isolated state where it was not safe for our monitoring agent to automatically fail over to the secondary replica without risking data loss, and manual intervention was required. We are actively working on a fix to resolve this issue.

By 12:04 PM PDT on April 21st, the outage was contained to the one affected Availability Zone and the degraded EBS cluster was stabilized. APIs were working well for all other

Availability Zones and additional volumes were no longer becoming "stuck". Our focus shifted to completing the recovery. Approximately 13% of the volumes in the Availability Zone

remained "stuck" and the EBS APIs were disabled in that one affected Availability Zone. The key priority became bringing additional storage capacity online to allow the "stuck" volumes

ot reuse the failed node until every data replica is successfully reidid not want to re-purpose this failed capacity until we were sure we apacity to replace that capacity in the duster. This required the timeapacity into the degraded EBS cluster. Second, because of the d the cluster in the step described above), the team had difficulty is to allow negotiation to occur with the newly-built servers without as the team had to navigate a number of issues as they worked ficant amounts of new capacity and working through the replication frected Availability Zone were restored by 12:30PM PDT on April pective of the attached EC2 instances because some were blocked ne and elect a new writible copy.

ess to the affected Availability Zone and restoring access to the raded EBS nodes to the EBS control plane and vice versa. This effort API access online to the impacted Availability Zone centered on instance of the EBS control plane, one we could keep partitioned to go. We rapidly developed throttlers that turned out to be too consethe moming of April 23rd, we worked on developing fine-grain s. Initial tests of traffic against the EBS control plane to the degraded negotiate which replica would be writable, to once again be usable Availability Zone.

nes in the Region. The recovery of the remaining 2.2% of affected the event as an extra precaution against data loss while the event and began processing batches through the night. At 12:30 PM PDT ected volumes. At this point, the team began forensics on the 0 PM PDT, the team began restrong these. Utilimately, 0.07% of the

"RDS"). RDS depends upon EBS for database and log storage, and as

hultiple Availability Zones ("multi-AZ"). Single-AZ database instances d if one of the EBS volumes it was relying on got "stuck". In the latively-bigger portion of the RDS population than the ases aggregate (/O capacity for database workloads under normal I the volume is restored. The percentage of "stuck" single-AZ d. The percentage of "stuck" single-AZ database instances in the hours, and the rest recovered throughout the weekend. Though we ity Zone had an underlying EBS storage volume that was not option to initiae point-in-time database restore operations.

ISC institute for SOFTWARE RESEARCH Carnegie Mellon University <u>School of Compu</u>ter Science

Cornerstones of Resilence

"[resilient is the] ability to sustain operations before, during, and after an unexpected disturbance"



- 1. Anticipation: know what to expect
- 2. Monitoring: know what to look for
- **3. Response:** know what to do
- **4. Learning:** know what just happened (e.g, postmortems)







ene with Janua Bubblin, Kriga Grislew, Julis Ellipaw, and Ton (Januard)

enaitmen for one of the large of orders study operations in the country. Some regimes have field, and on a bonne why. Steva lived; are off the shares while leaves of organizers well around the disk for three days due to stores.

The good zeros is the 43 and a study all about a changed is study have these brokeds. It is nevershow altergood to study a comparison here including the intercheding strength of study. These provides altergood and altergood are some complexely have of all or threat almostations and atoms requirement. Alters do before there and sponsed here, it has been provided by the 14 and 14 and

nett. As the only XXMs, denotes control Game (Mey, a program dispatch interview extinues of program), including some (datases nets control Games even impacted to distance Games) and dispatchesis. Batchards, Games (and the control on a summary systems, a diverse, and program (in the same of a supreme to a distances even. If the dispatch dispatchesis (for Games) and the solution of the same investment many sumports in our or to data and the same of a subject in the same of the solution of the same sumports in our or to data and there said in siding the structures manufactures manufactures.

Articipant solubilizes fields, the archite of classific of solutions shown in term in and (Salid) and the Marker Officiant Robinson solitor results on a pringities or advectiging considered, fishing indicating targeting of stabilised requires. In 610 Section 20 Nor and Fashing Technology and the opposities Constraints, the small Collidy constraints 20 Nor and parked for Vision (20 No Printtional Upposities, Constraints, the small Collidy constraints (20 North 20 No

Graph's operation of Garrel Day lorses by Krippe Kindowan, who has here with the program about from the free it study is sure-api. Six also works on other infrastructure projections of shells are freewal and the emission of sure and their date.

Making the document of Single obligation. The Educated States at the oblight regiment Making the document of Single obligation. The Annual Single obligation at the Soliton Single Obligation with Team Management for System Administration and The Practice of System and Network International Systems Single Obligation States States Single Obligation States Stat

TOM LIMONCELUS Since, you've preliably been involved in more GameDay emittees than anybody. What's the recoil important issues won've taken away from that?

"[...] get people throughout the organization to start building their anticipation muscles by **thinking about what might possibly go wrong.**"



Anticipation

These experiments form a cycle where developers **begin to anticipate what might possibly go wrong** during development, which adds to the overall resilience of the system.



Carnegie Mellon University School of Computer Science

Response: Etsy's Substitution Test

Developer runs command that brings down the site

- Grab another engineer who had no involvement in the incident
- Explain the context of the problem
- Fill developer in on the details known by the developer at the time
- Ask what they would do

Developer says they would run the same command almost every time

Identify the reasons for why this seemed the right decision at the time





instances for the effect of the second secon

cart prover falses, the loss to property for k to free to carpe it. More operators or relating to relations applicating on k hopes of hopesing largerises in falses, the where it free orders do to a solution and health care, and ears it is false any interfals to a farithe energy to high the industries task as arised on and health care, and ears it is being adapted to large cash. Web operators as well.

As the outp 2006, Annuary neutral Game (Eq. 4 program display) The investor obligate by particulation of any obligation of the obligation of the obligation of the obligation of a soft display display. Basis also, a classifier conclusion on a company's prime, and chemic, and and display display display. In a physical to a classifier conclusion of an employer's prime, and chemic, and people in the displayer of the displayer of

Steril Denova Nachan, S. Kaloba, D. and Anton C. Sanz A. San Chang, Andre Sen S. Alexin, and S. Sanz, Martin Sen and Senzia Senzia. The Anton Anton Senzia Senzi

pain kertensi. Gongle zuprinteri di "Laned-bay ternet-by Kolpa Kerlahana, arke her bert wilk the program obsert (her by their et started als pain ago. Ste also works on other infrastructure projects, most of which are found at

No principal of a new and level data. Balandating this decounters is a Complexificaper. Twee Lancescrift, where is a site establish engineers Multi-bases in a cyclew admented and controls, for charded out at AM Latic and has seelline from being ensure would be them Management for System Administration and The Practices of System and Network Internationalism.

TON LINOXCELLI Long, you've publishe bases involved in more GameDay exaction data amplitude. What's the more important insets you've taken away from that?

Some Example Google Issues

Terminate network in Sao Paulo for testing:

• Hidden dependency takes down links in Mexico which would have remained undiscovered without testing

Turn of data center to find that machines won't come back:

• Ran out of DHCP leases (for IP address allocation) when a large number of machines come back online unexpectedly.

Complexities are introduced as new capabilities are developed. [...] It gets progressively **harder to see where our dependencies are and what might lead to cascading failures.**

20





COMPLETE Case Study Learning to Embrace Fa

Prover wanty the holder, dropping waters and annetting is only sening at a 22to credit handling transactions (in one of the largest orders stud operations in the course, Some operate have field), and as not know why. Special and an off the charts while instea of organizers and provide the chart for three large toping we server.

to gate process to use or an evaluation of the strength of the

and: As the early AXMs, because is sensed from Line 2 surgears disspect to increase within a stype particular interface and a surgear serie regarded for shower lines and added agreed backs in Backshine at Construction and any symptyme and the analysis of the added agreed back and the structure and the structure particular structure and the structure

The comparison behavior of the second second

pair Relifies. Goight capitalists of 'GanedAir Issue by Aripa Reliability, who has been with the program advant Power by they it started air yours ago. She also works on other industricities praiects, must of which are focused on the started air yours ago.

represente of series and new states. Balanstate first discussion is a clough colleagues. Twee Lawner-PR, where a site stability requests Add Lawner is system administrator colles, he classed out at AM Labs and has wellen fair leads, most coldify Tame Management for System Administration and The Practice of System and Network.

W LINONCELLE long, you've periodity been involved in more GeneDay combine than anythody, but's the coust important instant work've taken away from that?

Netflix: Background

Started as a **DVD-by-mail business** because Reed Hastings was annoyed with Blockbuster late fees

Problem: when new movies come out, there's **only hundreds** of DVDs to **service multiple thousands of demand**

Stream movies instead of purchasing and mailing DVDs out to customers

Problem: must **purchase enough compute to handle peaks** (7pm+ weekends) vs valleys (noon, weekday)





Netflix: Cloud Computing

Significant deployment in Amazon Web Services in order to remain **elastic** in times of high and low load (first public, 100% w/o content delivery.)

Pushes code into production and modifies runtime configuration hundreds of times a day

Key metric: **availability**

a customer who can't watch a video because of a service outrage **might not be a customer for long.**



"Chaos Engineering" Basiri *et al., IEEE Software 2016*





Chaos Engineering: The History

Experimentation to **build confidence** around a system to **withstand turbulent conditions** in proven successful; today all Netflix

Netflix's Simian Army

- (the original) Chaos Monkey: Randomly terminates EC2 instances in production
- Chaos Kong:
 Simulates the failure of an entire EC2 region in AWS
- Latency Monkey:
 Injects latency to simula services react appropriately

Have Chaos Monkey crash development instances, too!

engineers design their services to handle instance failures as a matter of course.

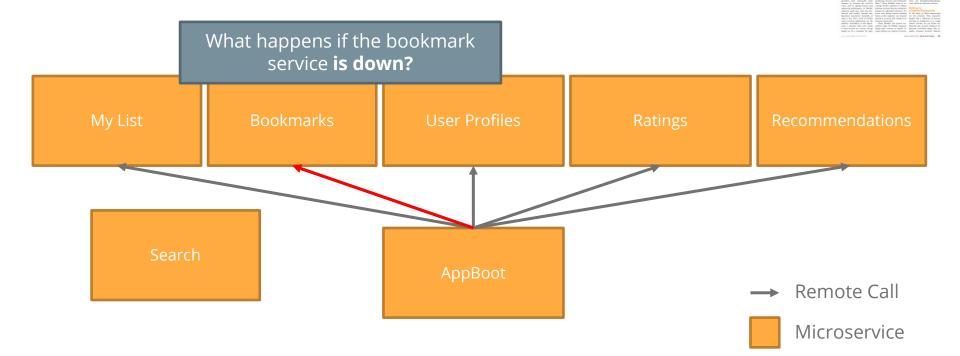




Chaos Engineer

Al Berr, Norte Betren, Norf & Norj, Lett Rodater, Lete Reservet, Faits Neytokik and Carep Rosettial Series & Modern archware based services are implemented

Netflix UI: AppBoot





Chaos

Principles of Chaos Engineering

- 1. Build a **hypothesis** around steady state behavior
- 2. Vary **real-world events** experimental events, crashes, etc.
- 3. Run **experiments** in production control group vs. experimental group draw However, **"works properly" is too vague** a basis for designing experiments.
- 4. Autornate experiments to ran continuously



<section-header><section-header><text><text><text><text><text><text><text>

Does everything seem to be working properly?

Are users complaining?

institute for Carnegie Mellon University SOFTWARE RESEARCH School of Computer Science

Graceful Degradation: Anticipating Failure

Allow the system to degrade in a way it's still usable

Fallbacks:

- Cache miss due to failure of cache;
- Go to the bookmarks service and use value at possible latency penalty

Personalized content, use a reasonable default instead:

- What happens if **recommendations** are unavailable?
- What happens if **bookmarks are unavailable?**

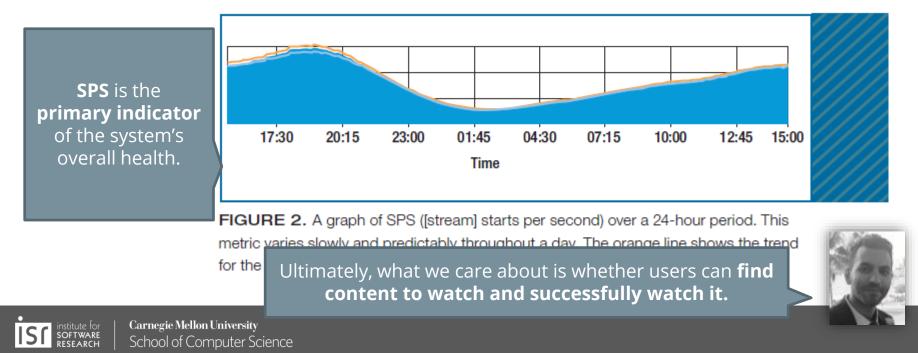
...default to starting videos at the beginning rather than providing a "resume from previous location" option.





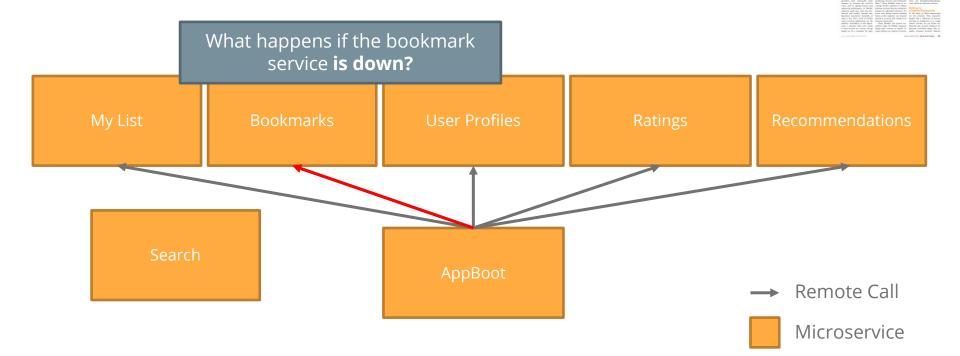
Steady State Behavior

Back to quality attributes: availability!



Chaos Engineerir

Netflix UI: AppBoot





Carnegie Mellon University School of Computer Science Chaos

AppBoot: Bookmarks Down Scenario (Imaginary)

SPS as core metric.

Experiment 1:

Outage of bookmarks service causes UI to fail to load, SPS decreases. Code fixed to hide bookmarks if call fails.

Experiment 2:

Outage of bookmarks service hides booksmarks on UI, SPS stays normal.



Exercise: Quality Attributes

- 1. What would a quality attribute be for an **e-commerce website** to characterize the stead-state behavior of the system?
- 2. What would a quality attribute be for an **advertisement platform** to characterize the stead-state behavior of the system?
- 3. What would a quality attribute be for an **admissions system** to characterize the stead-state behavior of the system?



Chaos

Making Hypotheses

No trivial hypotheses

- Overloading the system will increase the CPU, etc.
- Hypothesis should be made w.r.t overall system health metric

Monitor finer-grained metrics

- Monitor the CPU, other resources
- Indicators of degraded mode operation, etc.
- Use alerting to identify these issues to catch them early and anticipate





Varying Real-World Events

- 1. Clients send malformed requests
- 2. Servers may send malformed responses
- 3. Servers die
- 4. Hard disks fill up
- 5. Memory is exhausted
- 6. CPU is overloaded
- 7. Latencies spike
- 8. Load from clients can spike

A recent study reported that 92% of catastrophic system failures **resulted from incorrect handling of nonfatal errors.**





Sampling of Netflix's Candidate Faults

- 1. Terminate virtual machine instances
- 2. Inject latency into requests between different services
- 3. Fail requests between services
- 4. Fail an entire service
- 5. Make an entire Amazon **region unavailable**



Two Example Netflix Errors

- Server is overloaded and takes longer and longer to respond Clients requests are placed in a queue to be serviced Local queue becomes exhausted, run out of memory Client service crash
- Client makes a request to a server that uses a cache Error (transient) is returned to the client Server caches the error Future clients read the cached error value





Chaos Engineering as Continuous Process

Our system at Netflix changes **continuously**.

Because of these changes, our confidence in past experiments' results **decreases over time.**

Chaos Monkey **runs continuously** during weekdays, and we run Chaos Kong exercises monthly. (2016)

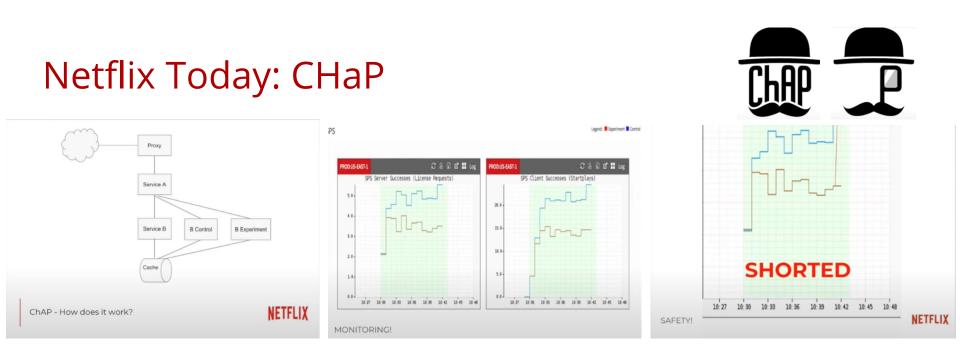


Chaos





Carnegie Mellon University School of Computer Science



Automatic experimentation and failure injection with FIT Automatic instrumentation of key performance metrics Automatic termination based on key metrics

Automatic experiment design with Monocle

reference: https://www.youtube.com/watch?v=3WRVgC8SiGc



How to run a Chaos Experiment

- 1. Define steady-state as some **measurable output of a system** that indicates normal behavior
- 2. Hypothesize that this steady state will continue in both the **control group** and **experimental group**
- 3. Introduce variables that **reflect real-world events** such server crashes, hard drives malfunctioning, and network connections being severed
- 4. Try to **disprove the hypothesis** by looking for a difference in steady state between the control group and the experimental group.

ST institute for Carnegie Mellon University RESEARCH School of Computer Science

