Papers we love too agenda

Thursday, May 20, 2021
7:00 PM to 9:00 PM PDT

PWL SF strictly adheres to the Code of Conduct (https://github.com/papers-we-love/papers-we-love/blob/master/CODE_OF_CONDUCT.md) set forth by all PWL charters.

https://www.meetup.com/papers-we-love-too/
https://github.com/papers-we-love/papers-we-love
https://www.youtube.com/user/PapersWeLove

Event link: https://www.meetup.com/papers-we-love-too/events/278269774

Roles

Host	Tech	Mini	Main
Aaron Goldman	Nikilesh	-	Yao Yue

Agenda

7:00 PM	Facilitator introduces Papers we love too Facilitator delivers announcements Tech lets attendees in from the waiting room Tech mutes attendees making noise during the talks Tech renames attendees with problematic names
7:05 PM	Facilitator introduces mini presenter 2 minutes
7:07 pm	Mini 15-20 minutes
7:27	Questions and answers 5-7 minutes
7:05 PM	break
7:10 PM	Facilitator introduces main presenter 2 minutes
7:15 PM	Main 35-45 minutes
8:00 PM	Questions and answers 5-7 minutes
8:10 PM	Guests plugs and open job adverts ● Announcements ○ Want to pick the papers you love join the organizing committee. ● Plugs

	○ Want to speak? Fill out https://goo.gl/forms/nt25JEk3bMkSEKQ73
8:30 PM	Meeting closed Guests welcome to stay on the zoom and network

Mini

- Presenter bio
  -
- Paper
  -

Main

- Presenter bio
  - Yao Yue
  - Bio: Yao Yue is an engineer and manager working at Twitter Platform. She has been working on distributed cache since 2010, with extensive experience with popular OSS projects such as Memcached and Redis. She designed and implemented a modular open-sourced cache framework called Pelikan (more at pelikan.io). Since 2017, she has started and managed the Infrastructure Performance and Optimization team. Her team work on infrastructure performance and capacity monitoring, optimizing systems configurations and utilization, cross-service tracing an insight at scale, and advancing software design to take advantage of new hardware technologies.
- Slides
  - https://docs.google.com/presentation/d/1eFY0oK_NPszfovCbxCK9d4Ho9niDV-q0WhTM0203tFo
  - https://github.com/thinkingfish/misc/blob/master/talks/Tail-at-Scale.pdf

- Paper
  - The Tail at Scale
    - Software techniques that tolerate latency variability are vital to building responsive large-scale Web services.
  - By Jeffrey Dean and Luiz André Barroso
  - https://cacm.acm.org/magazines/2013/2/160173-the-tail-at-scale/fulltext
  - Systems that respond to user actions quickly (within 100ms) feel more fluid and natural to users than those that take longer. Improvements in Internet connectivity and the rise of warehouse-scale computing systems have enabled Web services that provide fluid responsiveness while consulting multi-terabyte datasets spanning thousands of servers; for example, the Google search system updates query results interactively as the user types, predicting the most likely query based on the prefix typed so far, performing the search and showing the results within a few tens of milliseconds. Emerging augmented-reality devices (such as the Google Glass prototype) will need associated Web services with even greater responsiveness in order to guarantee seamless interactivity. It is challenging for service providers to keep the tail of latency distribution short for

interactive services as the size and complexity of the system scales up or as overall use increases. Temporary high-latency episodes (unimportant in moderate-size systems) may come to dominate overall service performance at large scale. Just as fault-tolerant computing aims to create a reliable whole out of less-reliable parts, large online services need to create a predictably responsive whole out of less-predictable parts; we refer to such systems as “latency tail-tolerant,” or simply “tail-tolerant.” Here, we outline some common causes for high-latency episodes in large online services and describe techniques that reduce their severity or mitigate their effect on whole-system performance. In many cases, tail-tolerant techniques can take advantage of resources already deployed to achieve fault-tolerance, resulting in low additional overhead. We explore how these techniques allow system utilization to be driven higher without lengthening the latency tail, thus avoiding wasteful overprovisioning.

Topic: Yao Yue on "The Tail at Scale"
Time: May 20, 2021 07:00 PM Pacific Time (US and Canada)

Join Zoom Meeting
https://us02web.zoom.us/j/87899574055?pwd=M1R5TVJDN3lodjVDV3A1K0pjYmw5Zz09

Meeting ID: 878 9957 4055
Passcode: 405410
One tap mobile
+16699009128,,87899574055#,,,,*405410# US (San Jose)
+13462487799,,87899574055#,,,,*405410# US (Houston)

Dial by your location
+1 669 900 9128 US (San Jose)
+1 346 248 7799 US (Houston)
+1 253 215 8782 US (Tacoma)
+1 646 558 8656 US (New York)
+1 301 715 8592 US (Washington DC)
+1 312 626 6799 US (Chicago)
Meeting ID: 878 9957 4055
Passcode: 405410
Find your local number: https://us02web.zoom.us/u/k4j3UIps0