Scaling Challenges at Two Sigma

Mark Astley
Head of Reliability Engineering

Disclaimer

This document is being distributed for informational and educational purposes only and is not an offer to sell or the solicitation of an offer to buy any securities or other instruments. The information contained herein is not intended to provide, and should not be relied upon for, investment advice. The views expressed herein are not necessarily the views of Two Sigma Investments, LP or any of its affiliates (collectively, “Two Sigma”). Such views reflect the assumptions of the author(s) of the document and are subject to change without notice. The document may employ data derived from third-party sources. No representation is made by Two Sigma as to the accuracy of such information and the use of such information in no way implies an endorsement of the source of such information or its validity.

The copyrights and/or trademarks in some of the images, logos or other material used herein may be owned by entities other than Two Sigma. If so, such copyrights and/or trademarks are most likely owned by the entity that created the material and are used purely for identification and comment as fair use under international copyright and/or trademark laws. Use of such image, copyright or trademark does not imply any association with such organization (or endorsement of such organization) by Two Sigma, nor vice versa.

Two Sigma Overview

- Diversified financial advisor (~$58B AUM)
- Data-driven strategies
- Sophisticated analytics and modeling
- Founding member of Mass Open Cloud
  - Mentoring and course projects
  - Hardware donations
  - Special projects – storage, security, etc.

Two Sigma 10 Years Ago

- Markets
- Vendors
- Curated sources
- 100’s of datasets
- 10’s GB/day

- Model Development
- Backtest
- ~10K cores
- ~50TB memory

Significant infrastructure investment required

High bar for competition, disruption difficult

Scaling and Agility Challenges Today

Data Bazaars

Google Dataset Search

- Markets
- Vendors
- Curated sources
- Direct sources
- Unstructured data
- 1000's of datasets
- 10's TB/day

Data → **Analytics And Modeling**

- Model Development
- Backtest
- Machine Learning
- ~500 GPUs
- ~100K cores
- ~500TB memory

Public Cloud

*aaS:
Platform
Infrastructure
Storage
ML

Infrastructure no longer a bar for competition

A Brief List of Challenges

- It’s a distributed world...
  - 10 years ago, a single machine for modeling was sufficient
  - Today - distributed systems rule (spark, neural nets, data flow, etc.)
- It’s a distributed world (part 2)...
  - 10 years ago, data could fit on a single machine or database
  - Today - data locality issues are important, need to move compute to data
- Casting a wider net for data...
  - Everyone can get curated data without much investment
  - Uncurated, sometimes self-acquired, data has high value
  - Data provenance will become critical
- Hybrid environments will dominate...
  - No single, isolated technology platform is likely to suffice
  - The most capable platforms will span multiple environments
  - New security processes will need to be developed

Where Mass Open Cloud Fits In

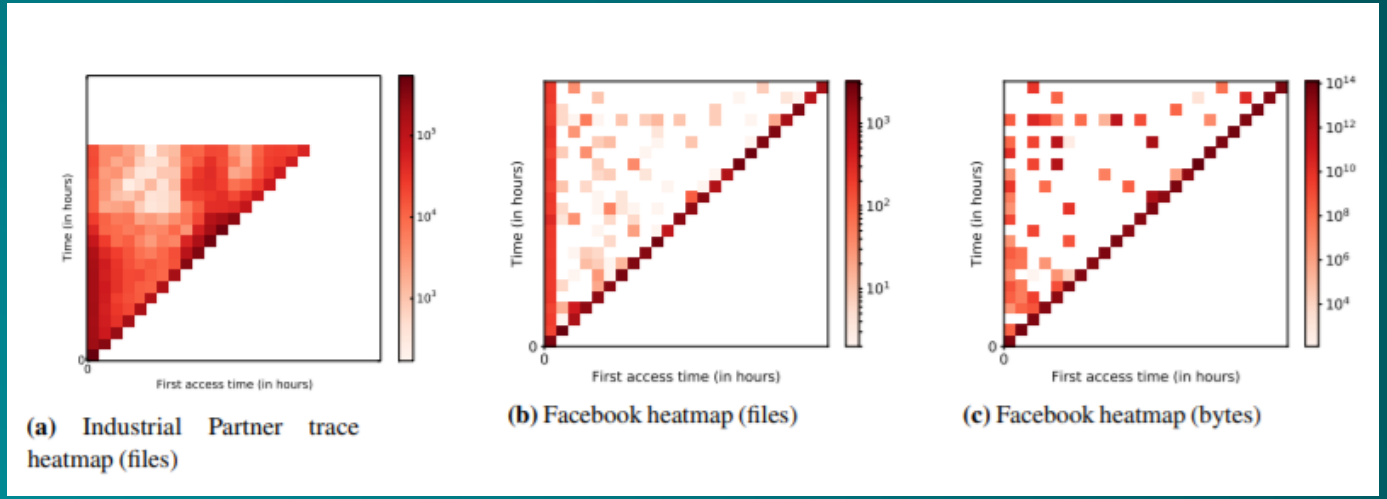
- Diversify our bets on technology and innovation
  - “Non-correlated” ideas for next generation infrastructure.
  - What can be built with best of breed?
  - What if we pursue a different agenda than the public clouds?
- Scale models for Scaling Problems
  - Build hybrid and/or scaled infrastructure
  - Build data “bazaar” and other data access models
  - Experiment with new secure compute capabilities

MOC Participation

- Student Projects
  - Functions as a Service on OpenStack
  - Serverless Performance with Containers: Docker Swarm on OpenStack
    - OpenStack Summit 2017, Rohit Kumar
  - Location-Aware Computing

- Research Collaboration and Publication
  - On-site visits, TS -> MOC and vice versa
  - Larry R. on site to advise

Collaborative Research - Improved Storage Caching



(a) Industrial Partner trace heatmap (files)

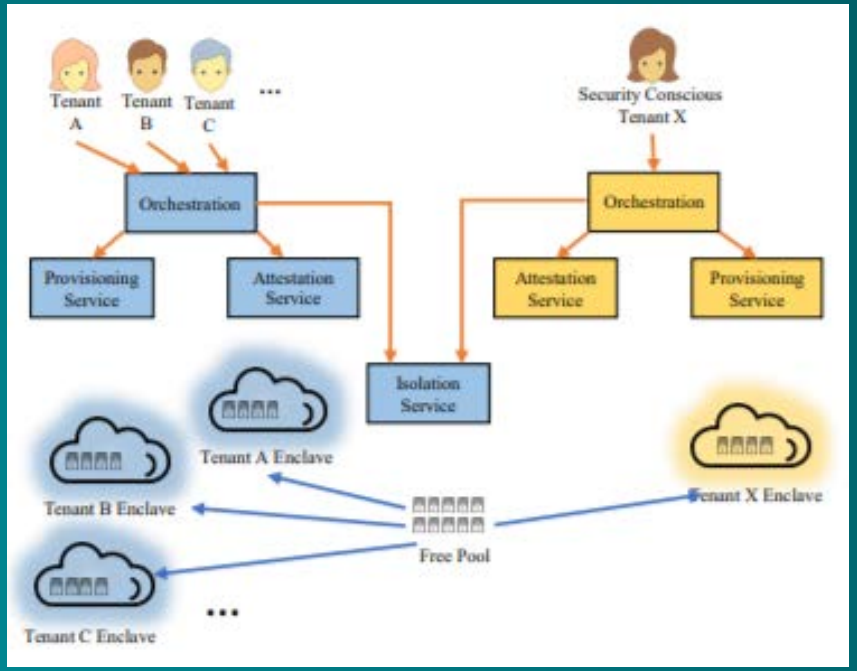
(b) Facebook heatmap (files)

(c) Facebook heatmap (bytes)

D3N: A multi-layer cache for improving big-data applications' performance in data centers with imbalanced networks

- *Kaynar et al*, USENIX ATC '18 Poster Session, Eurosys '19 (submitted)

Collaborative Research - Security on Bare Metal Clouds



A Secure Cloud with Minimal Provider Trust

- *Mosayyebzadeh et al*, HotCloud ‘18

Future Goals

- Partner better with MOC participants
  - Opportunities to collaborate around our shared future
- Expand interactions beyond usual partners
  - What can we learn from MOC “customers”?
  - What can we learn from other academic partners?
- Explore Machine Learning Collaborations
  - Will there be a machine learning “bazaar” to accompany data?
- Increase student engagement and Two Sigma interactions

Thank You!