LONDON
INTERNATIONAL
SOFTWARE DEVELOPMENT
CONFERENCE 2015

goto; conference

Keeping your Cloud Footprint in Check

Coburn Watson

goto; conference

Click ‘engage’ to rate session.

Rate **12** sessions to get the supercool GOTO reward

Join the conversation #gotoldn

@coburnw

- Cloud Performance and Reliability @ Netflix
  - Reduce TTD and TTR
  - Build innovative performance analysis tooling
  - Optimize usage of AWS Cloud
  - Steer global user traffic and support failover
  - Inject Chaos into production environment
  - Drive operational best practice adoption

NETFLIX

goto; conference

- 67M+ Subscribers
- > 50 countries
- > 3 billion hours of video streamed monthly
- Huge cloud footprint
- Homegrown CDN
- Strong Originals slate

NETFLIX OSS

goto; conference

- Strong focus on open source efforts
- https://netflix.github.io/

Atlas

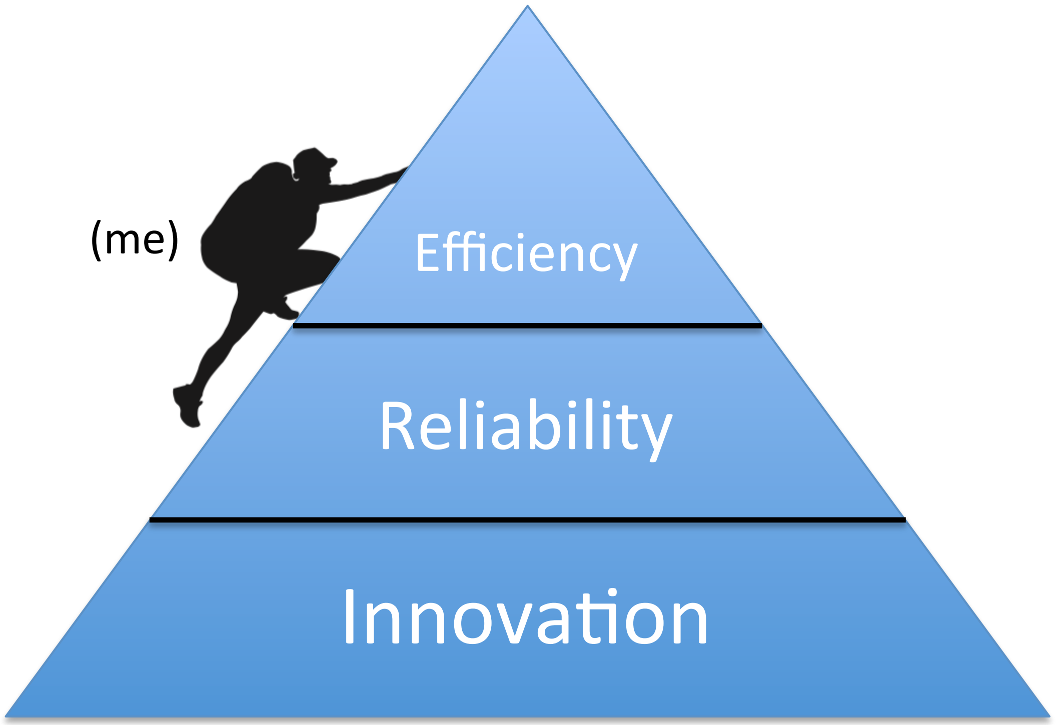
ASGARD

ICE

Vector

HYSTRIX
DEFEND YOUR APP

Our Priorities

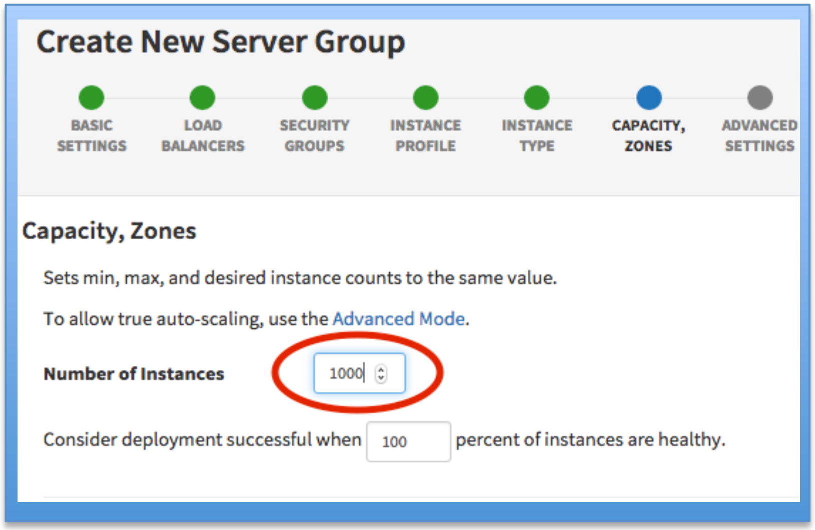

(me)
Efficiency
Reliability
Innovation

goto; conference

Cost of Innovation and Reliability

Maximize Innovation

- Capacity On-Demand
- Commit-to-Cloud in minutes
- Single Production Account (~ 350 µservices)
- Burst into on-demand, cover with reservation purchases



Create New Server Group

Capacity, Zones

Sets min, max, and desired instance counts to the same value.

To allow true auto-scaling, use the Advanced Mode.

Number of Instances 1000

Consider deployment successful when 100 percent of instances are healthy.

Cost of Reliability

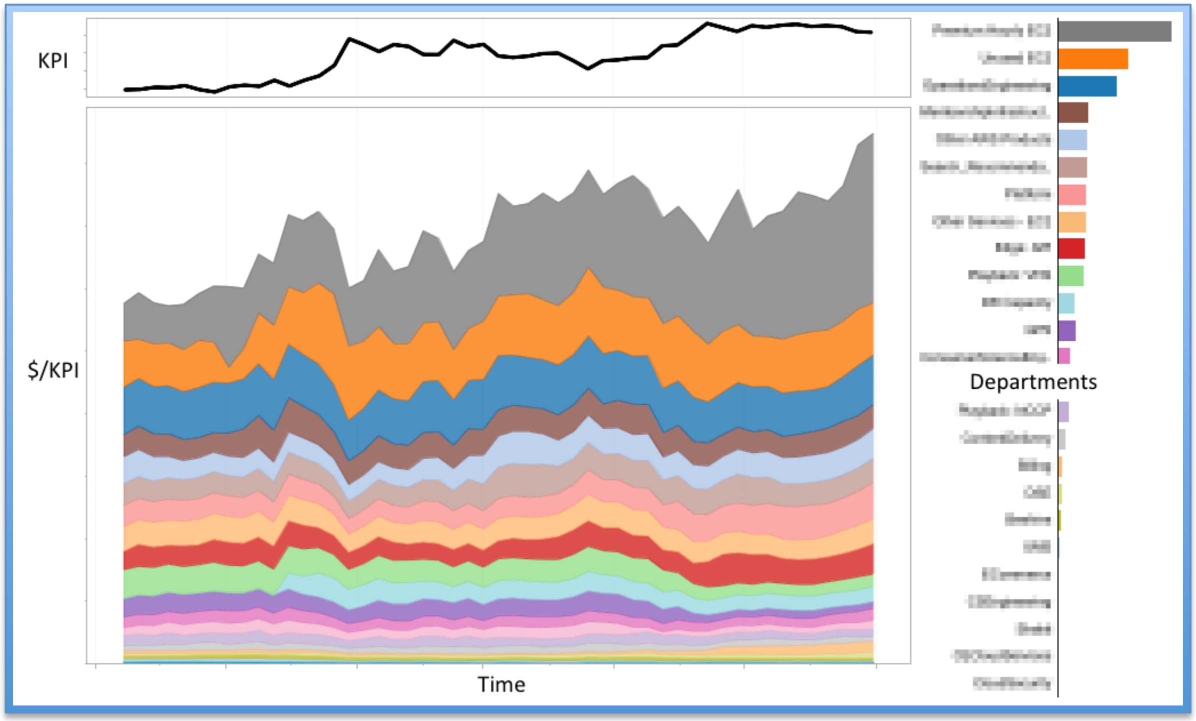
- Red-Black push model
- Over-provision for redundancy in AWS Region
- Global redundancy through failover
- Purchase “Heavy” AWS EC2reservations to secure capacity

goto; conference

Efficiency

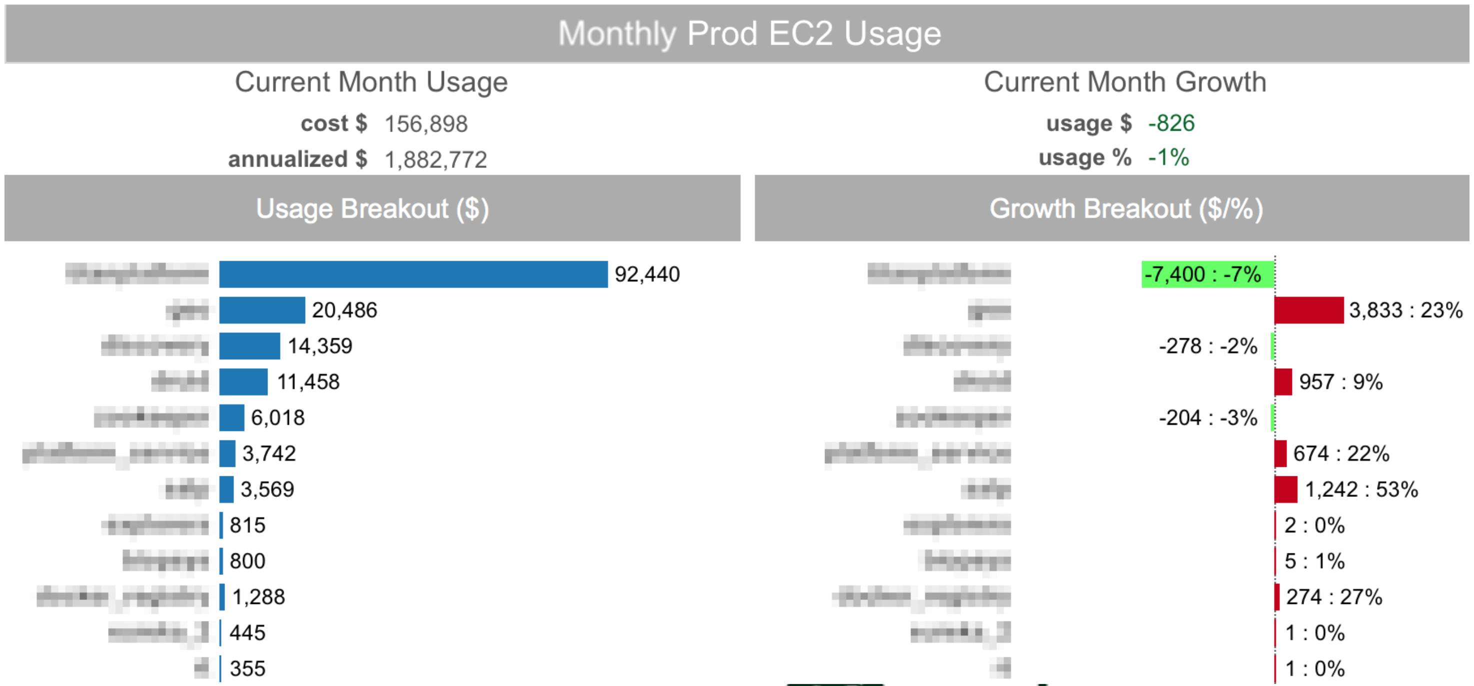
Efficiency Goals

- Have them and track them!



Monitoring Costs

- ICE: Open Source AWS Cost Monitoring Utility
- Internal Cost Reporting pushed to first-level managers



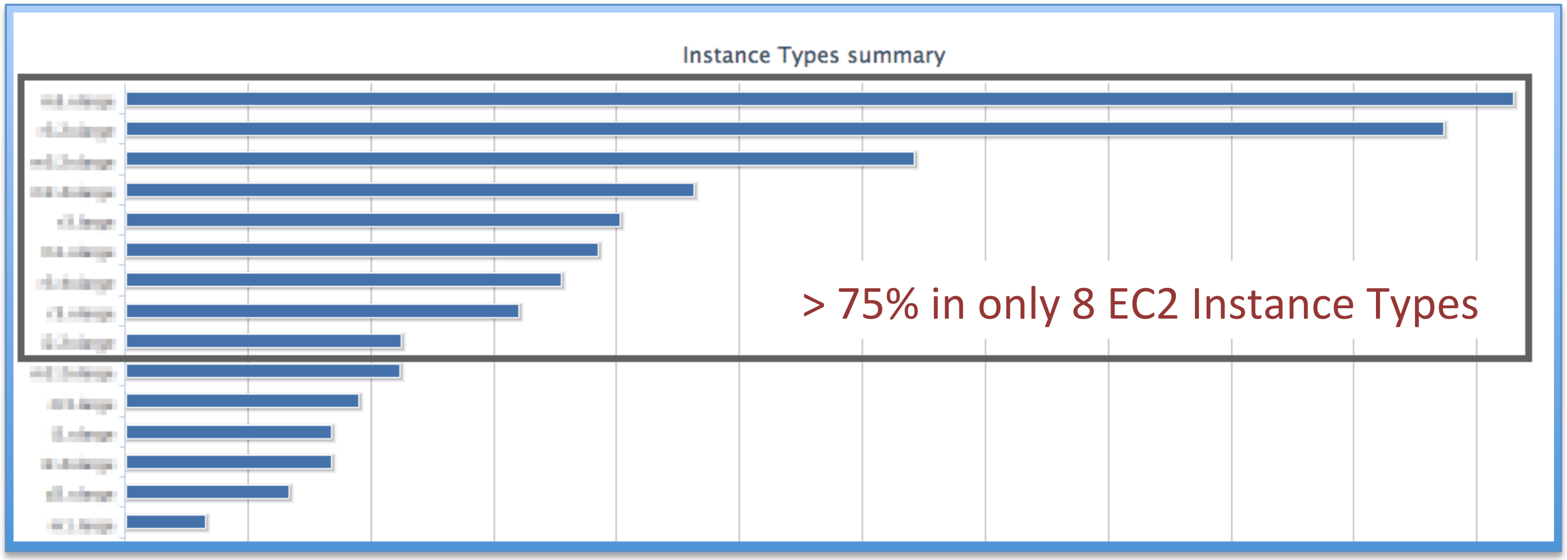
Monthly Prod EC2 Usage

Current Month Usage		Current Month Growth	
cost $	156,898	usage $	-826
annualized $	1,882,772	usage %	-1%

Usage Breakout ($)		Growth Breakout ($/%)	
[redacted]	92,440	[redacted]	-7,400 : -7%
[redacted]	20,486	[redacted]	3,833 : 23%
[redacted]	14,359	[redacted]	-278 : -2%
[redacted]	11,458	[redacted]	957 : 9%
[redacted]	6,018	[redacted]	-204 : -3%
[redacted]	3,742	[redacted]	674 : 22%
[redacted]	3,569	[redacted]	1,242 : 53%
[redacted]	815	[redacted]	2 : 0%
[redacted]	800	[redacted]	5 : 1%
[redacted]	1,288	[redacted]	274 : 27%
[redacted]	445	[redacted]	1 : 0%
[redacted]	355	[redacted]	1 : 0%

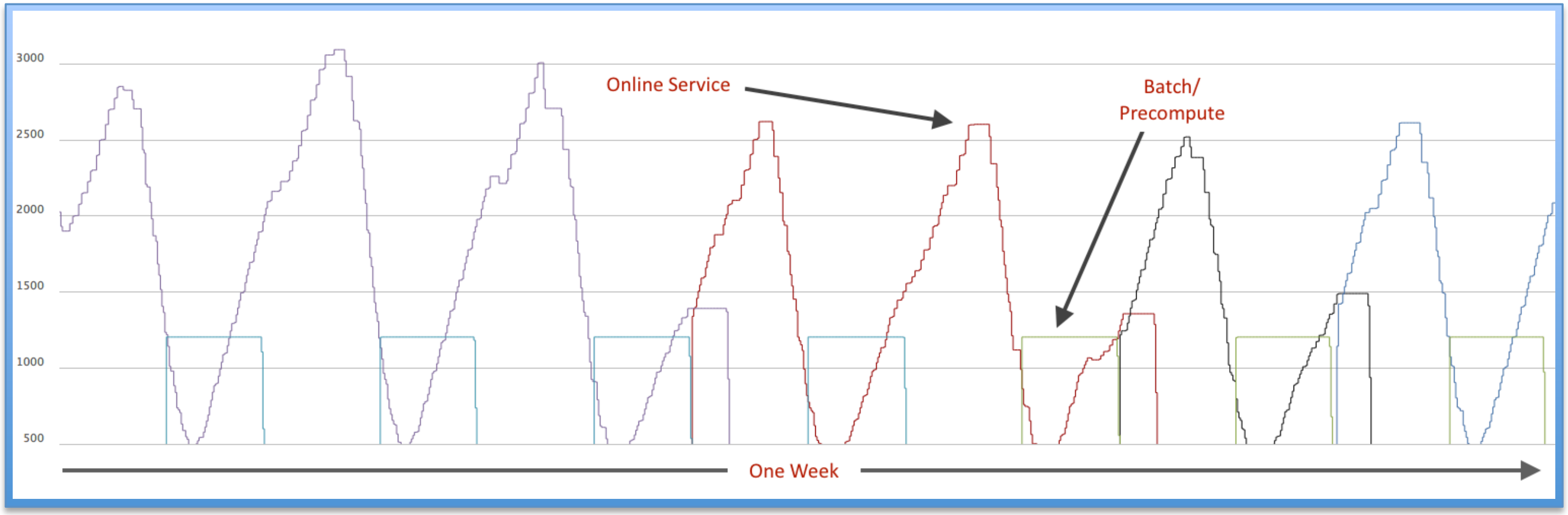
Maximize Sharing

- Single Production Account
- Fewer/Larger Pools
- Maximize Shared Capacity



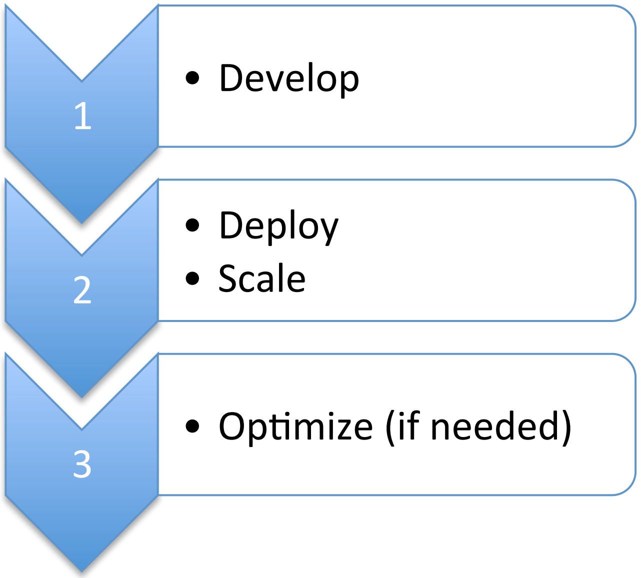
Encourage Borrowing

- All accounts are linked at a billing level
- Large troughs of unused capacity exist (Autoscaling)
- Interruptible workloads for internal “Spot”



Optimization

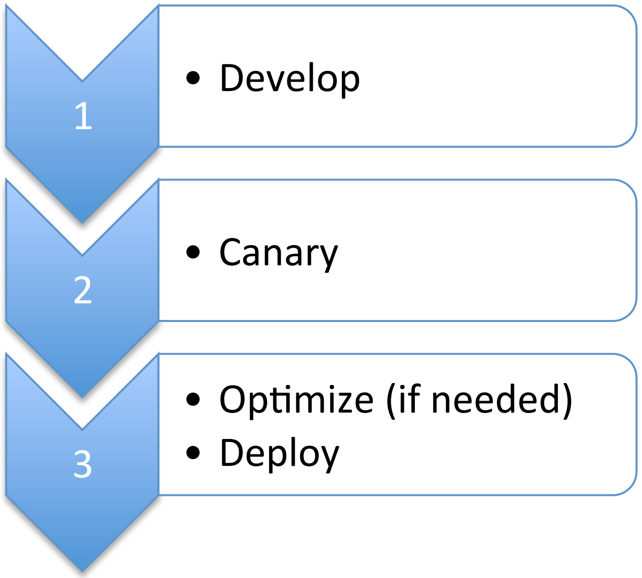
- Direct Consultation for “Big Fish”
- Tooling for Everyone


1
- Develop

2
- Deploy
- Scale

3
- Optimize (if needed)

New Services or Features


1
- Develop

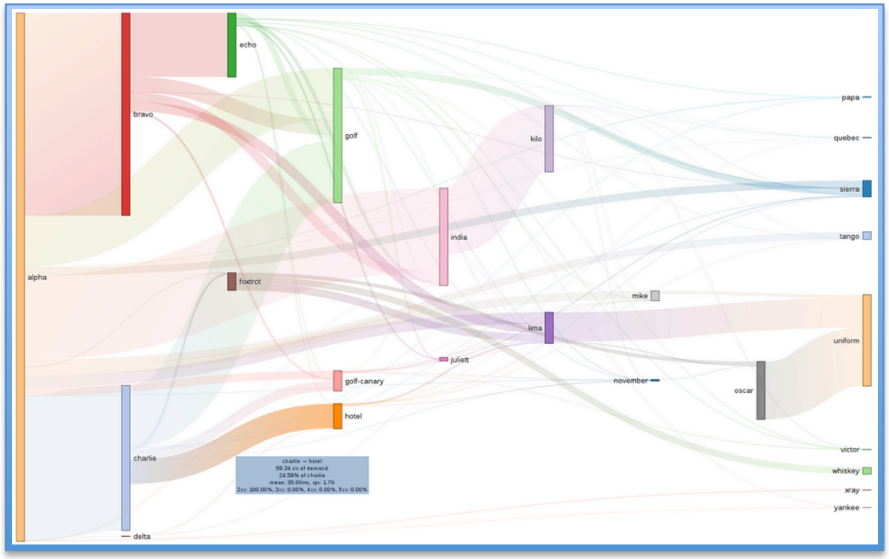
2
- Canary

3
- Optimize (if needed)
- Deploy

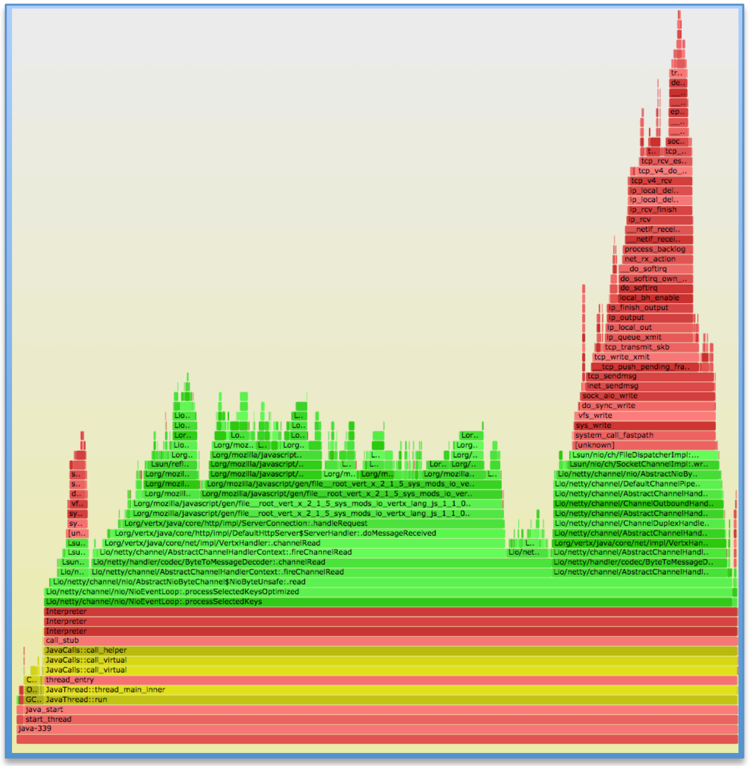
Ongoing Service Development

Improving Stack Observability

- Too big for commercial tools
- Patch key middleware where necessary



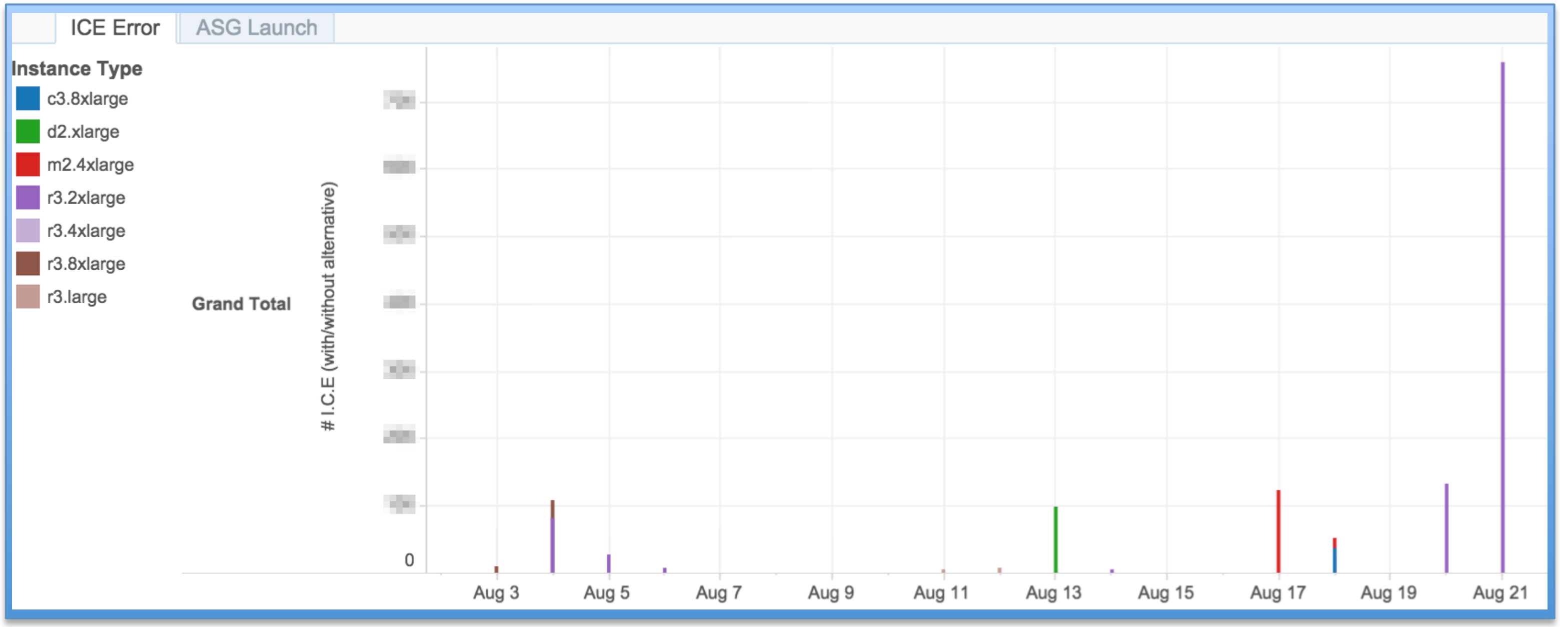
Transaction Tracing with Resource Demand



Mixed-Mode JVM CPU Flame Graph

Monitor Capacity Shortfalls

- Constrain On-Demand charges
- Identify/alert on significant capacity provisioning events



Data Points

- Internal Borrowing
  - Encoding consumed 135k cross-account EC2 Instance hours June 2015 (> ~ $200k/monthly savings)
  - Data Platform (Hadoop, etc.) saves > $1MM/year

Summary

- Target your Innovation:Efficiency ratio
- Push cost context to the team level
- Embrace the elasticity of the Cloud

goto; conference

Please

Remember to rate this session

Thank you!

Join the conversation #gotoldn

goto; conference

Thanks !