

Performance Engineering for SaaS

Kusuma Seshavarapu



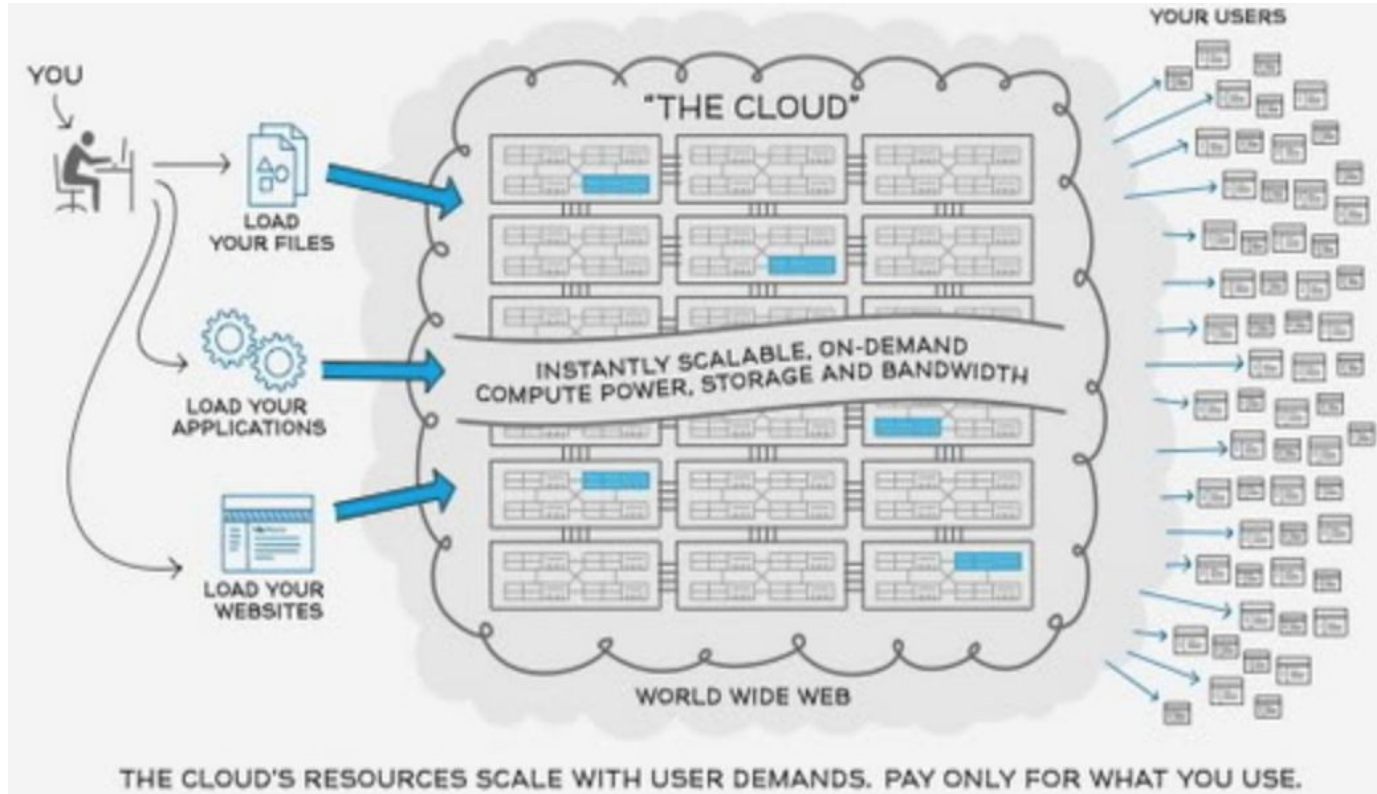
What does performance mean?



HIGH
AVAILABILITY



What does Cloud Offer Us?



Do we need to worry about Performance in SaaS??





Performance Matters the Most in SaaS

SaaS providers are bound by SLAs

- Availability (99.9)
- Performance(High Apdex score)
- Scalability (on-demand)
- RPO: Recovery Point Objective
- RTO: Recovery Time Objective
- Other QoS(Security, compliance)

SLA Breach results in



Performance matters -Revenue



found that a **2** second slowdown = **4.3%** reduction in revenue/user



stated that a **400** millisecond delay → **0.59%** fewer searches/users



Noticed that users who experience the fastest page load times view **50%** more pages/visits than users experiencing the slowest page load times



reduced page load times from ~7 seconds to ~2 seconds, leading to a **7-12%** increase in revenue

98% of organizations say a single hour of downtime costs over \$100,000



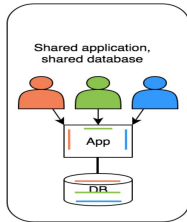
Performance matters - Customer Satisfaction

Outages or Latency could lead to

- Lost opportunity
- Shaken customer loyalty
- Damaged reputation

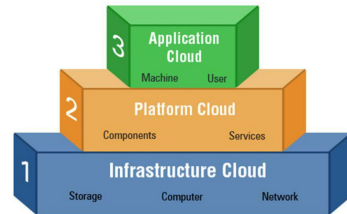
What makes the SLA compliance hard?

- Performance of a SaaS application depends on the performance of PaaS which in turn depends on IaaS



- Multi-tenancy- load of one tenant impacts others

- Failure is Assured - Lot of moving parts
- Hard to debug and test





Architectural Principles

- Build Highly Cohesive and Loosely Coupled System--think Microservices
- Stateless systems- scale out
- Autoscale & load balance each component and each layer-elasticity
- Avoid Single Point Of Failures
- Build for Failure--redundancy in all layers,
- Automate Everything
- Measure & Monitor all the resources



Performance Engineering Workflow

Define KPIs: Expected Load, Response time targets

Identify the right testing tools: ex: JMeter, Gatling

Identify the test types: Stress tests, Endurance tests, Reliability tests

Assimilate performance testing into the general development process: CI/CD

Collect & Monitor System Metrics: Integrate with APM

Test and Tune individual components/services

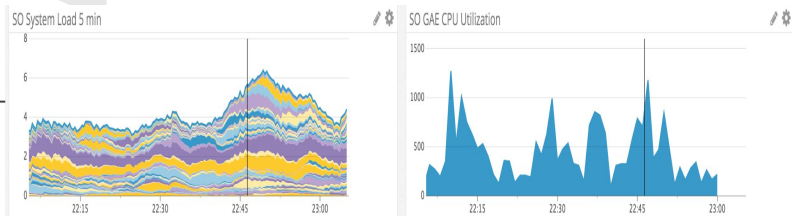
Establish the behavior of the application and the footprint it has in the cloud env

Estimate the system resources/cost for the performance modeling volume

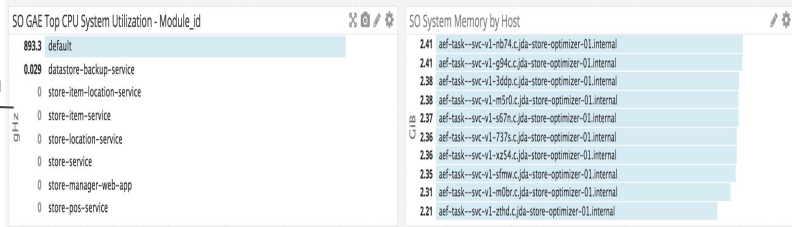
Come up with thresholds for resource throttling

Performance Vitals-Services dashboard

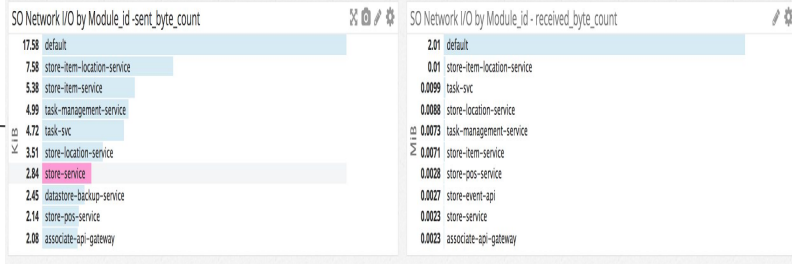
CPU



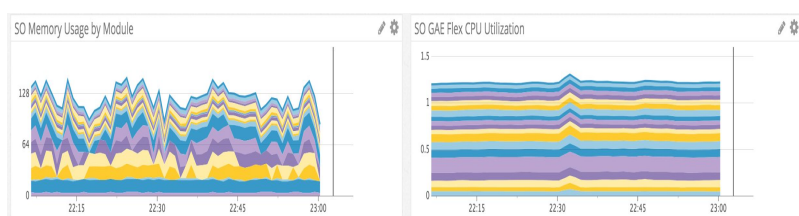
System Load



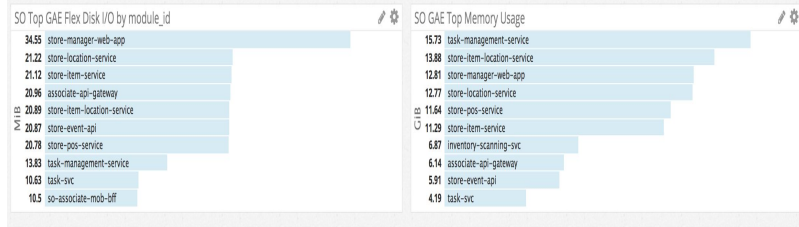
Network ingress/egress



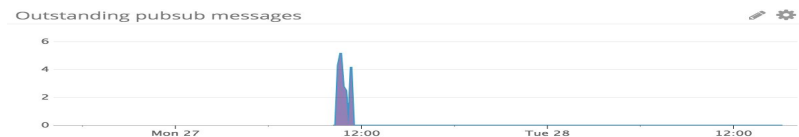
Memory



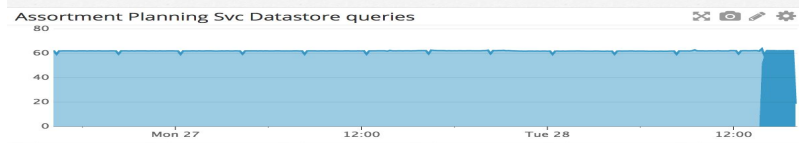
Disk IO



Msgs



Queries























Uptime checks

Uptime Checks



 Filter...

CHECKS	VIRGINIA	OREGON	IOWA	BELGIUM	SINGAPORE	SAO PAULO	POLICIES	ACTIONS
assortment-planning-app/_ah/health	✓	✓	✓	✓	✓	✓		
assortment-planning-svc/_ah/health	✓	✓	✓	✓	✓	✓		
assortment-scoring-app/_ah/health	✓	✓	✓	✓	✓	✓		
assortment-scoring-engine/_ah/health	✓	✓	✓	✓	✓	✓		
assortment-scoring-service/_ah/health	✓	✓	✓	✓	✓	✓		
authorization-service/_ah/health	✓	✓	✓	✓	✓	✓		
cloudstorage-bucket-service/_ah/health	✓	✓	✓	✓	✓	✓		
configuration-service/_ah/health	✓	✓	✓	✓	✓	✓		
datastore-backup-service/_ah/health	✓	✓	✓	✓	✓	✓		



Trace

Trace list

[+ ANALYZE RESULTS](#)

UNDO ZOOM

AUTO RELOAD

1 hour 4 hours 12 hours **1 day** 3 days 1 week 1 month Custom

Request filter

HTTP method

All

HTTP status

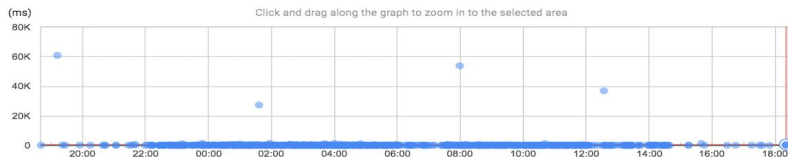
All

Service

All services

Version

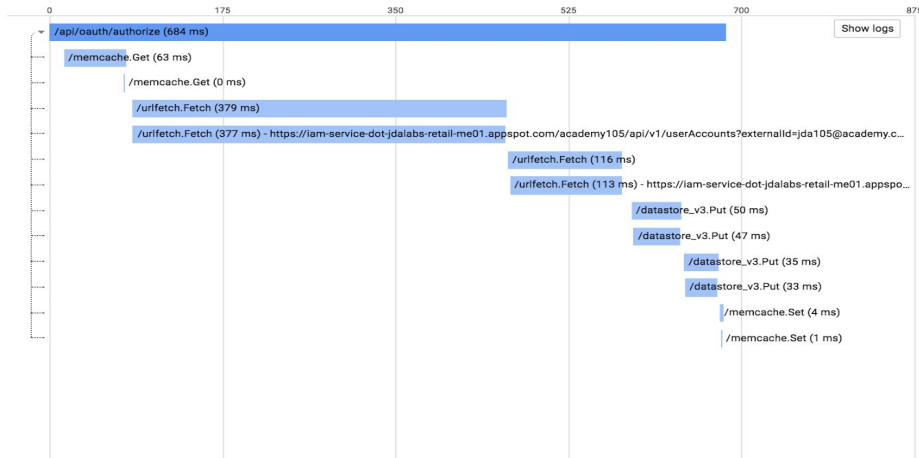
All versions



Latency	HTTP Method	URI	Analysis Report	Time
116 ms	GET	/api/oauth/authorize		6:23 PM (11 minutes ago)
684 ms	GET	/api/oauth/authorize		6:21 PM (14 minutes ago)
332 ms	GET	/api/oauth/authorize		6:20 PM (14 minutes ago)
112 ms	GET	/api/oauth/authorize		6:20 PM (14 minutes ago)
104 ms	GET	/api/oauth/authorize		5:49 PM (45 minutes ago)

Rows per page: 5 1 - 5 of 1000 < >

Timeline



@0 ms

/api/oauth/authorize

Summary









Name	RPCs	Total Duration (ms)
/api/oauth/authorize	1	684
/datastore_v3.Put	4	165
/memcache.Get	2	63
/memcache.Set	2	5
/urifetch.Fetch	4	985

Details

Timestamp	2017-11-28 (18:21:01.054)
Traced time	647 ms
Untraced time	37 ms
Log	View
Report	View
Service	authorization-service
Version	authorization-service:v103.405824249073065560
HTTP method	GET



Performance Monitors & Alerts

<input type="checkbox"/>	STATUS	NAME	GROUP	TRIGGERED ↓
<input type="checkbox"/>	WARN	launch-pad-web-app: 500 Response Count	response_code:502,version_id:v27	 9¼h
<input type="checkbox"/>	NO DATA	store-event-api: HTTP Error Response Alert	response_code:502,version_id:v23	 11h
<input type="checkbox"/>	WARN	store-event-api: Average Latency Alert	response_code:502,version_id:v23	 11h
<input type="checkbox"/>	ALERT	so-associate-mob-bff: Network I/O Alert	version_id:v6	 13h
<input type="checkbox"/>	ALERT	authorization-service: Network I/O alert	version_id:v102	 13h
<input type="checkbox"/>	ALERT	so-associate-mob-bff: Disk I/O Alert	version_id:v6	 13h
<input type="checkbox"/>	ALERT	item-service: Network I/O alert	version_id:v13	 14h
<input type="checkbox"/>	WARN	store-admin-web-app-api: Latency Alert	module_id:store-admin-web-app...	 17h

Questions??

