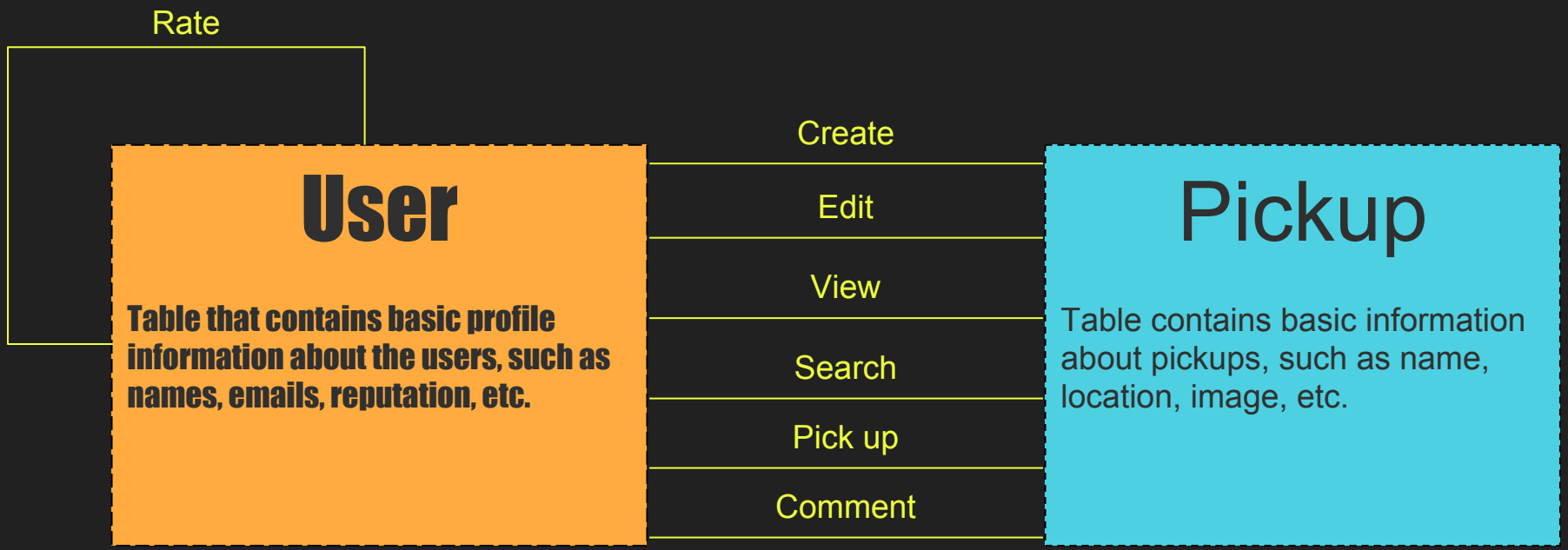
Pickachu

Tai-Yi Wu, Beiqi Guan,
Yao-Jen Chang, Pengcheng Pan

Application architecture



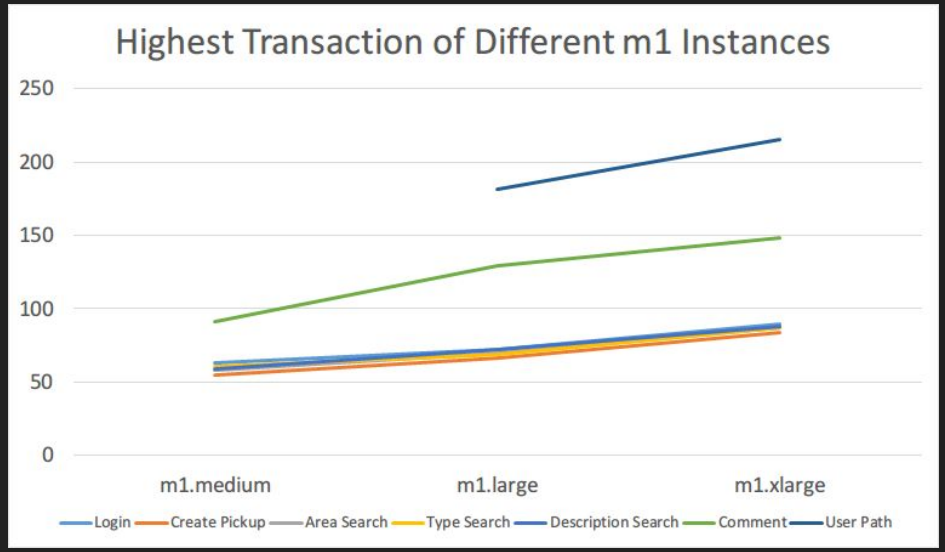
Unit Load Testing and Critical User Path

6 Unit Load Testing:

1. Login/Logout functionality
2. Create pickups after user login
3. Search in certain distance after user login
4. Search specific type of pickups after user login
5. Search keyword in pickups' description after user login
6. Search seller and give evaluation on certain pickup after user login

Critical User Path Load Testing: walk through a registered user path

Scaling performance results - Vertical



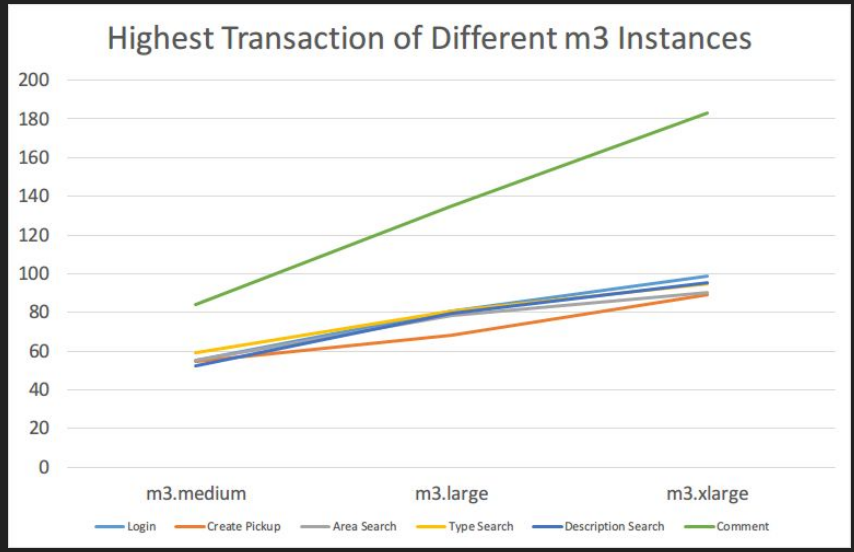
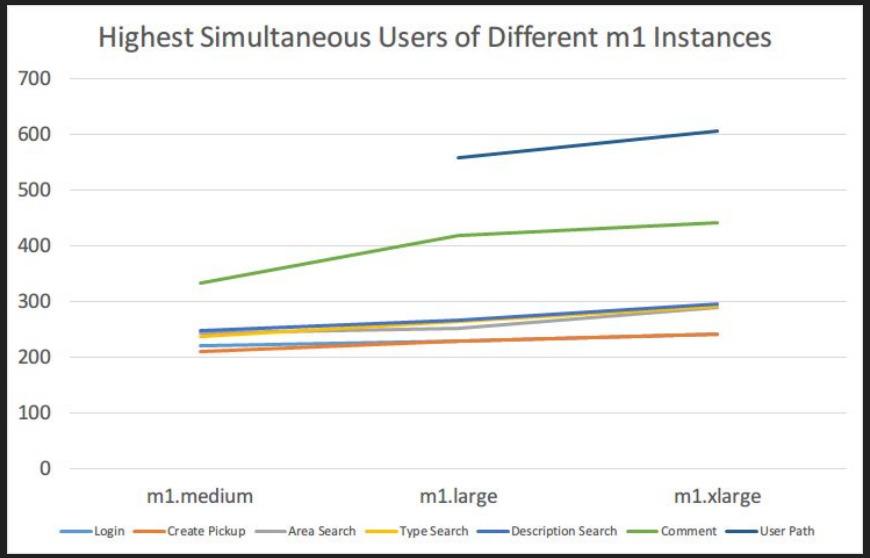


Chart generated from the data that we get.

Scaling performance results - Vertical



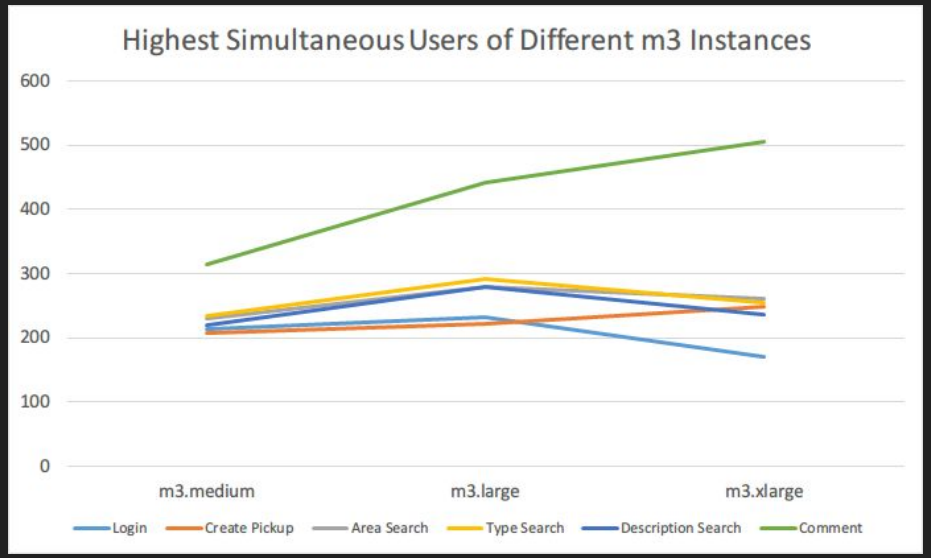
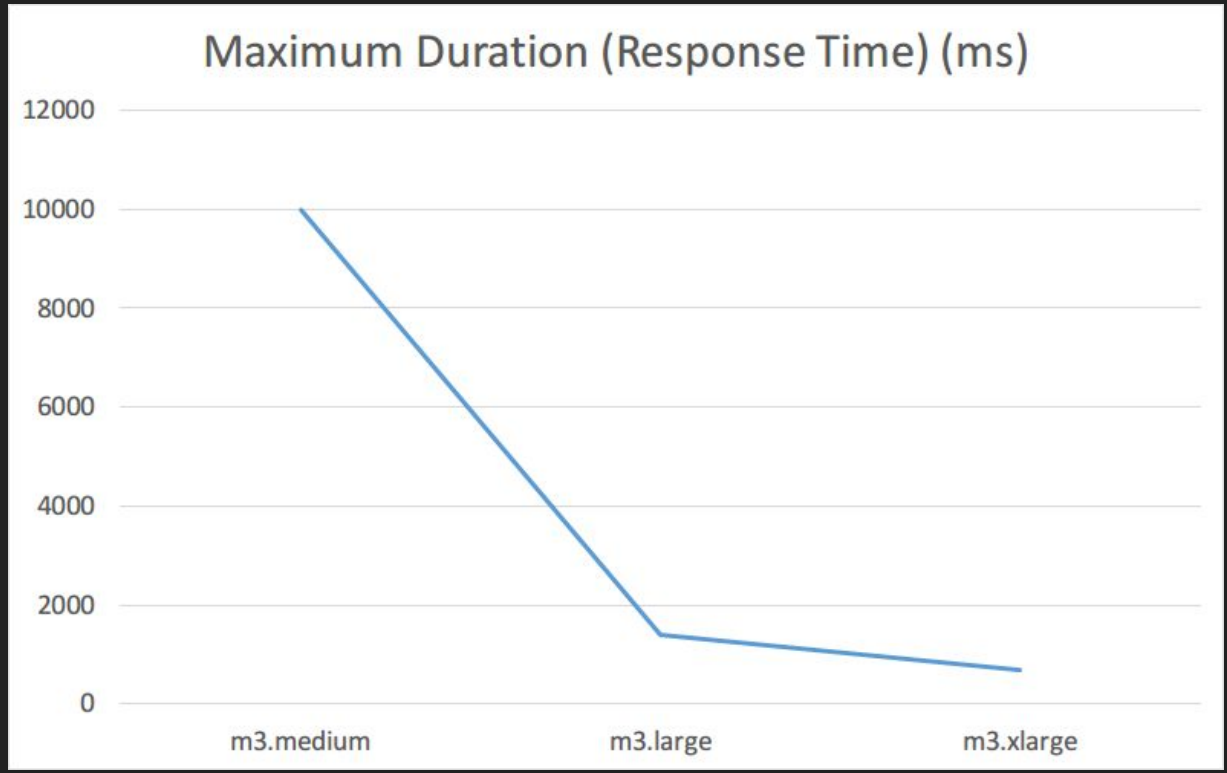


Chart generated from the data that we get.

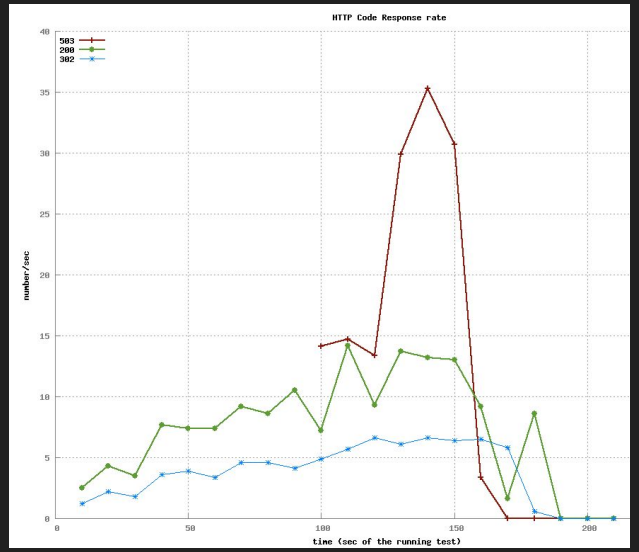
Scaling performance results - Vertical

- Use 1 unit load testing as example.
- From 10,000ms to 700ms! (93% faster)

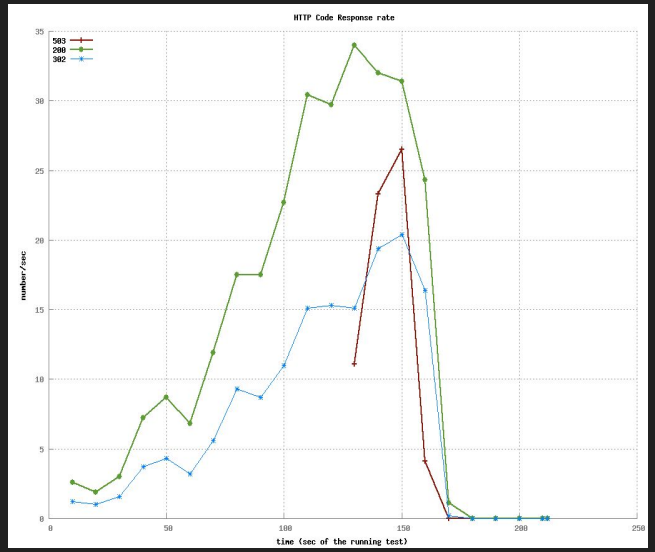


Scaling performance results - Vertical

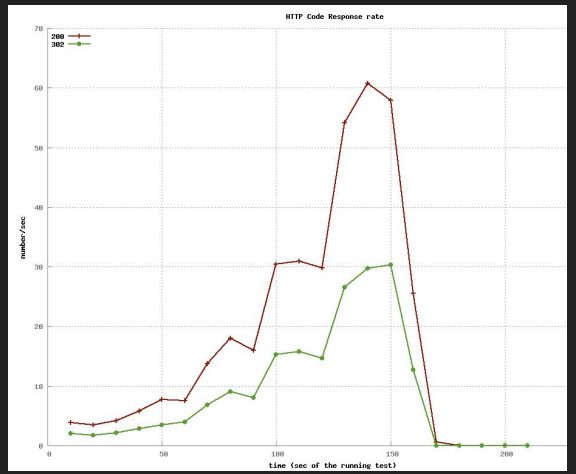
Http return code status of m3.medium, m3.large, m3.xlarge (1 unit load testing)



503 200 302
m3.medium



503 200 302
m3.large



200 302
m3.xlarge

Summary of vertical scaling

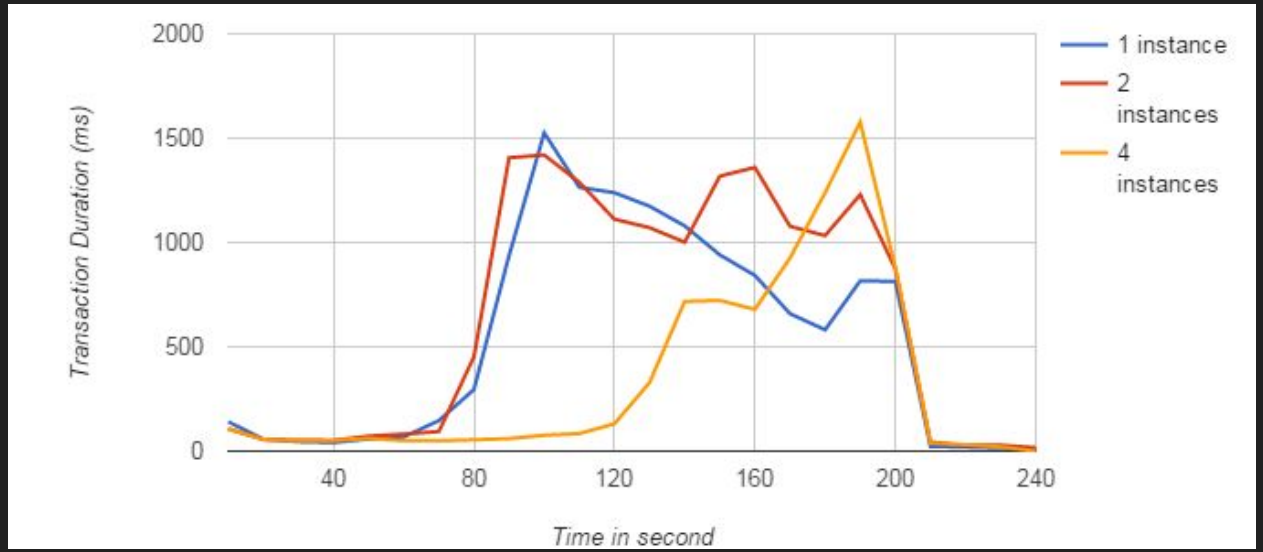
- Test m1.medium, m1.large, m1.xlarge, m3.medium, m3.large, m3.xlarge
- As we choose better instance:
  - Figher tansaction rate
  - Faster response time
  - Higher simultaneous ssers
  - Handle more HTTP requests

	Highest Transaction (Page/Sec)	Maximum Mean Transaction Duration Time (msec)
m3.medium	55.1	10,000
m3.large	78.3	1400
m3.xlarge	90.3	700

Sample comparision from report

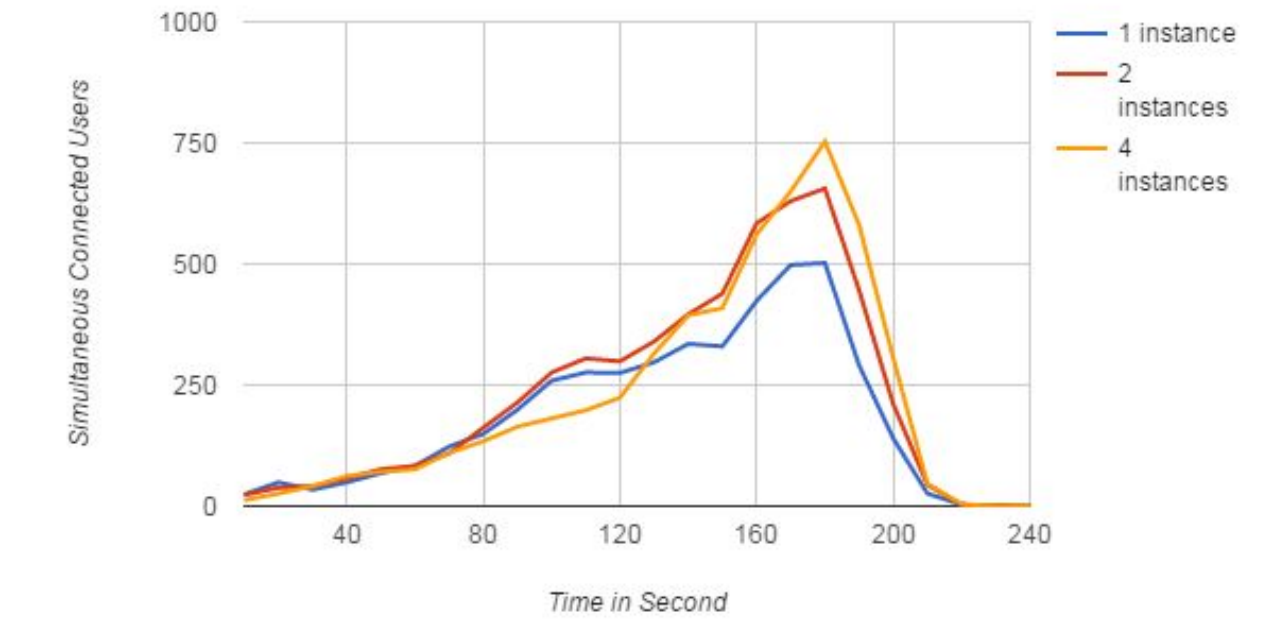
Scaling performance results - Horizontal

Phase	1	2	3	4	5	6
Users/sec	2	4	8	12	16	32



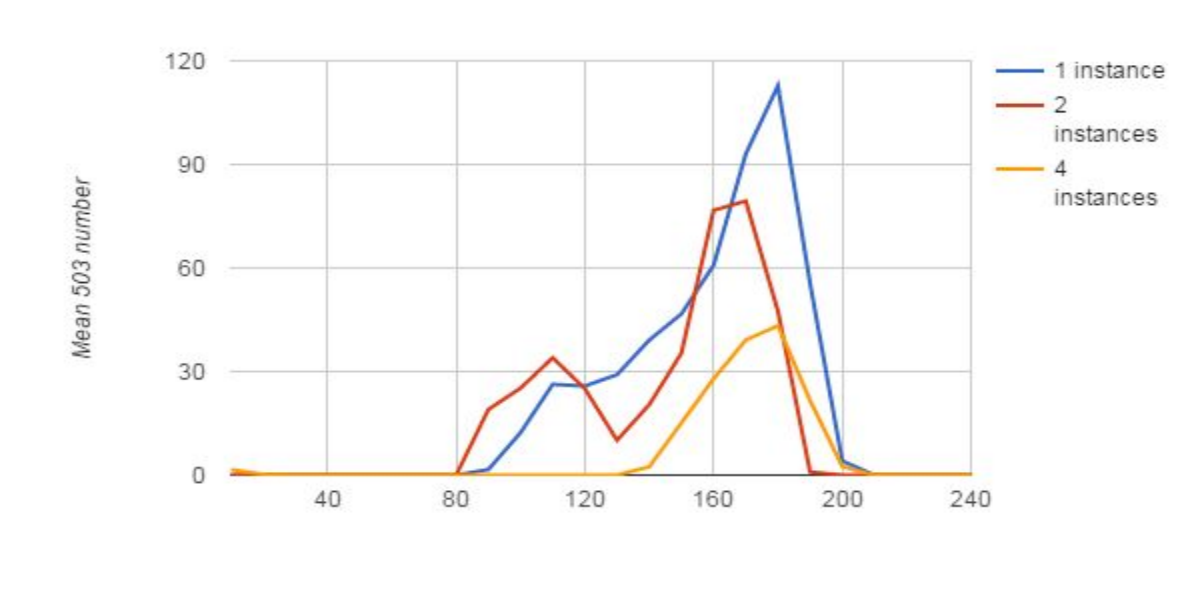
Mean Transaction Duration (Response Time) over Time

Scaling performance results - Horizontal



Simultaneous Connected Users over Time Series

Scaling performance results - Horizontal



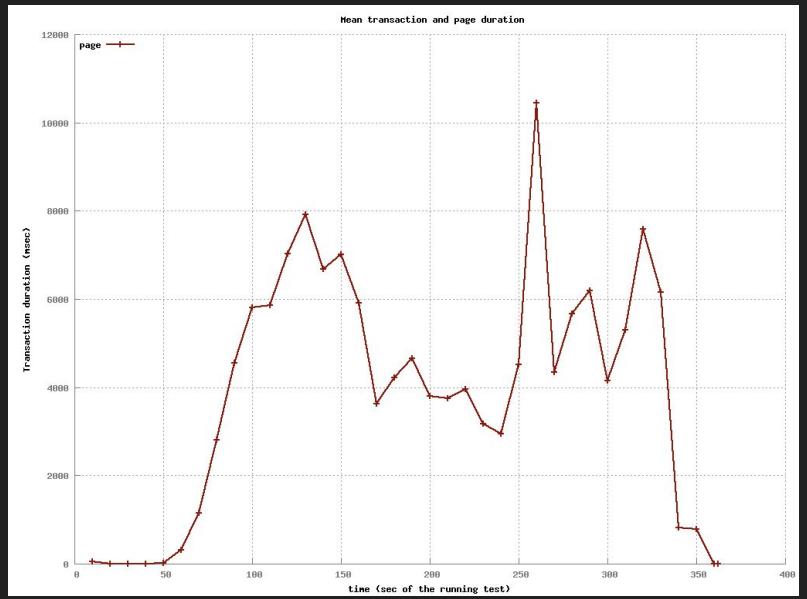
503 Server unavailable over Time

Summary of horizontal scaling

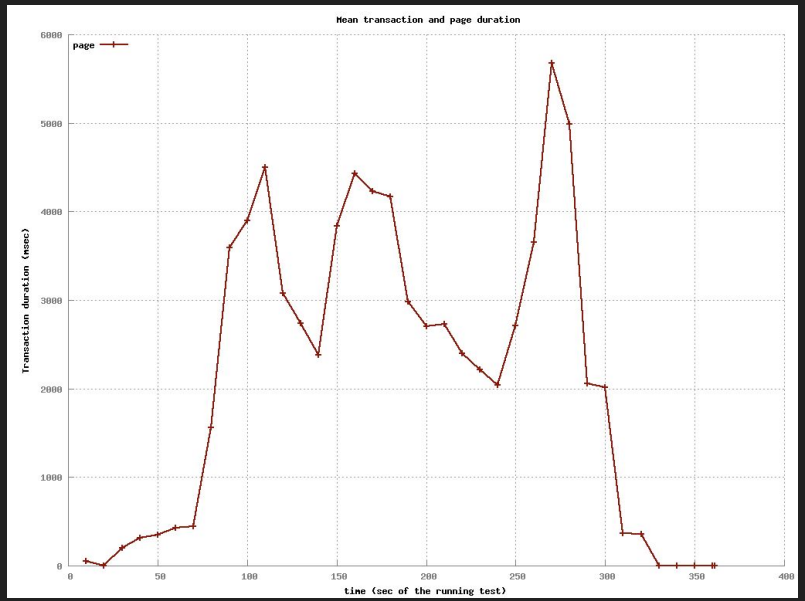
	Arrival rate to reach server error 503 (Users/Sec)	Maximum Simultaneous users without server error 503	Time to reach Maximum Mean Transaction Duration Time(sec)	Maximum Mean Transaction Duration Time (msec)
1 Instance	8	149	90	1525.4
2 Instances	8	162	100	1419.34
4 Instances	16	394	190	1577.06

Optimization - Improvement with File-based cache

Phase	1	2	3	4	5	6	7	8
User/sec	2	4	8	12	16	32	48	64

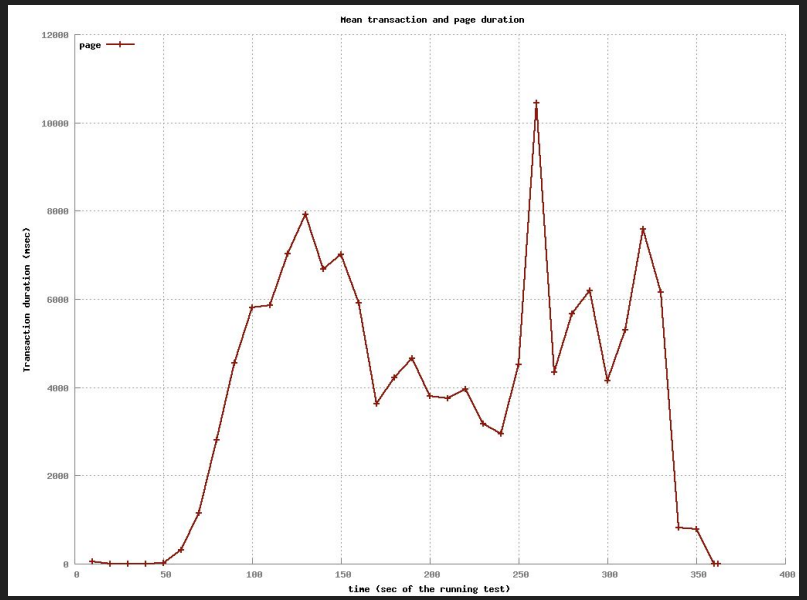


Master branch without cache file

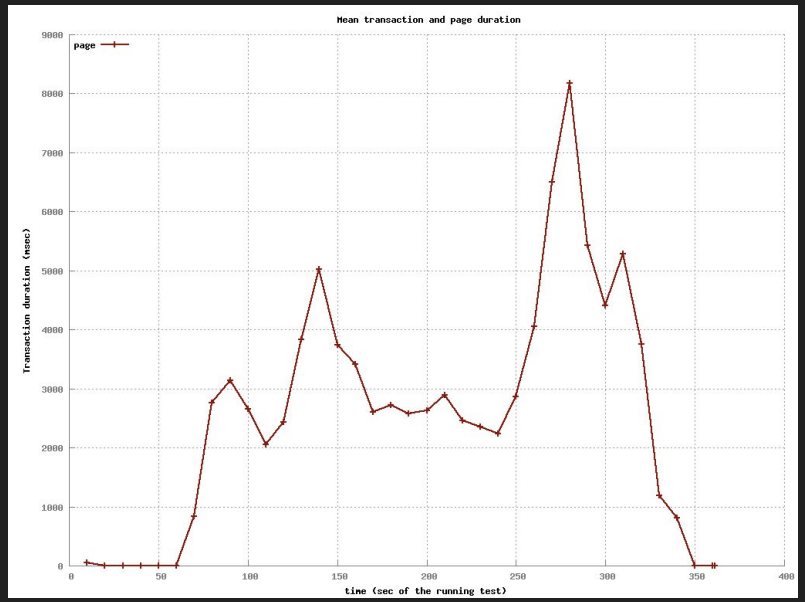


Master branch with cache file

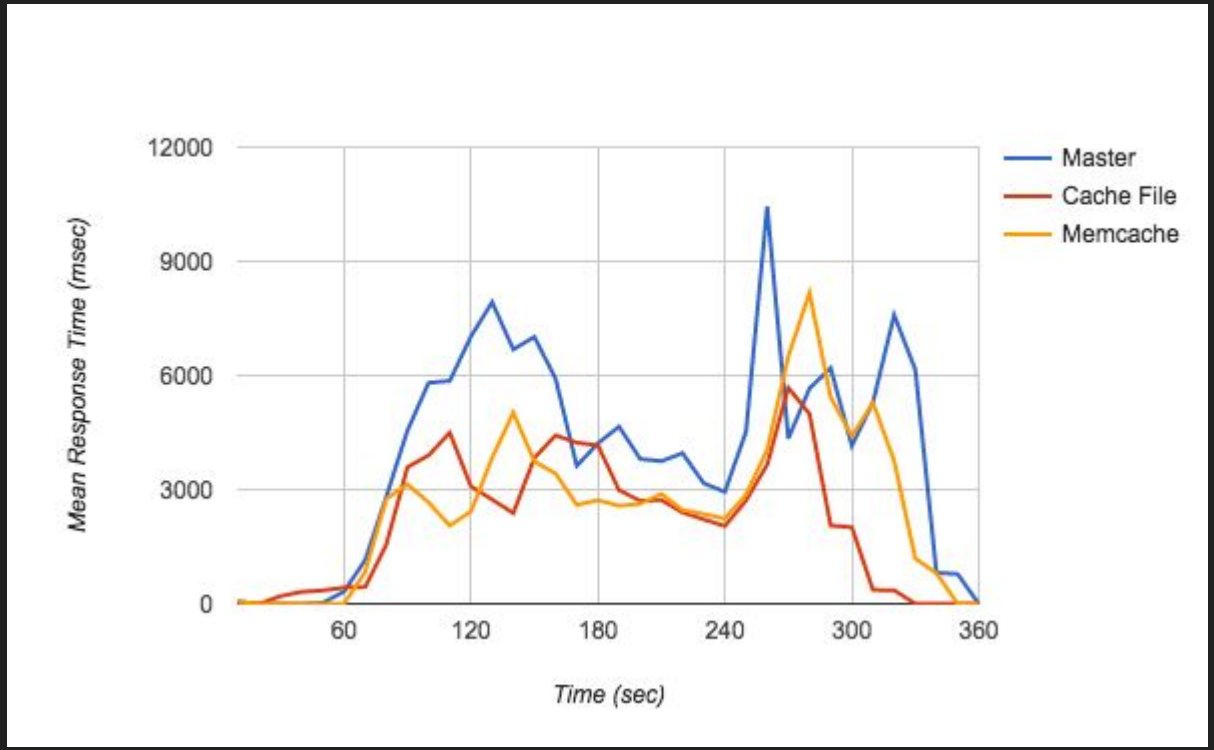
Optimization - Improvement with Memcache



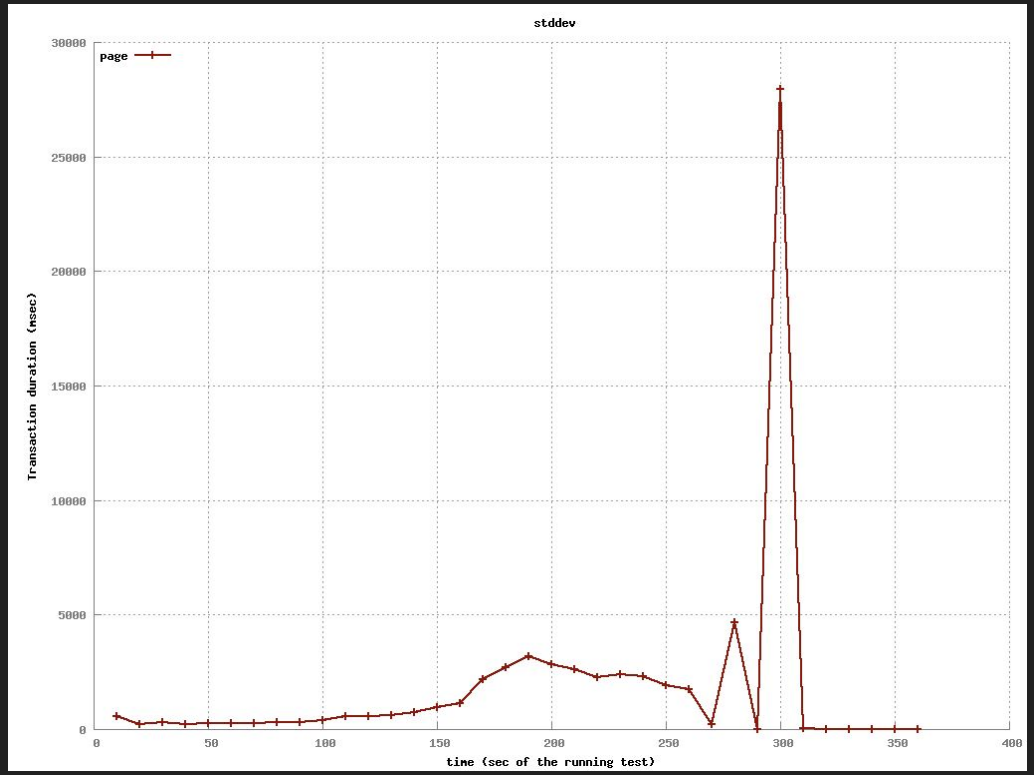
Master branch without memcache



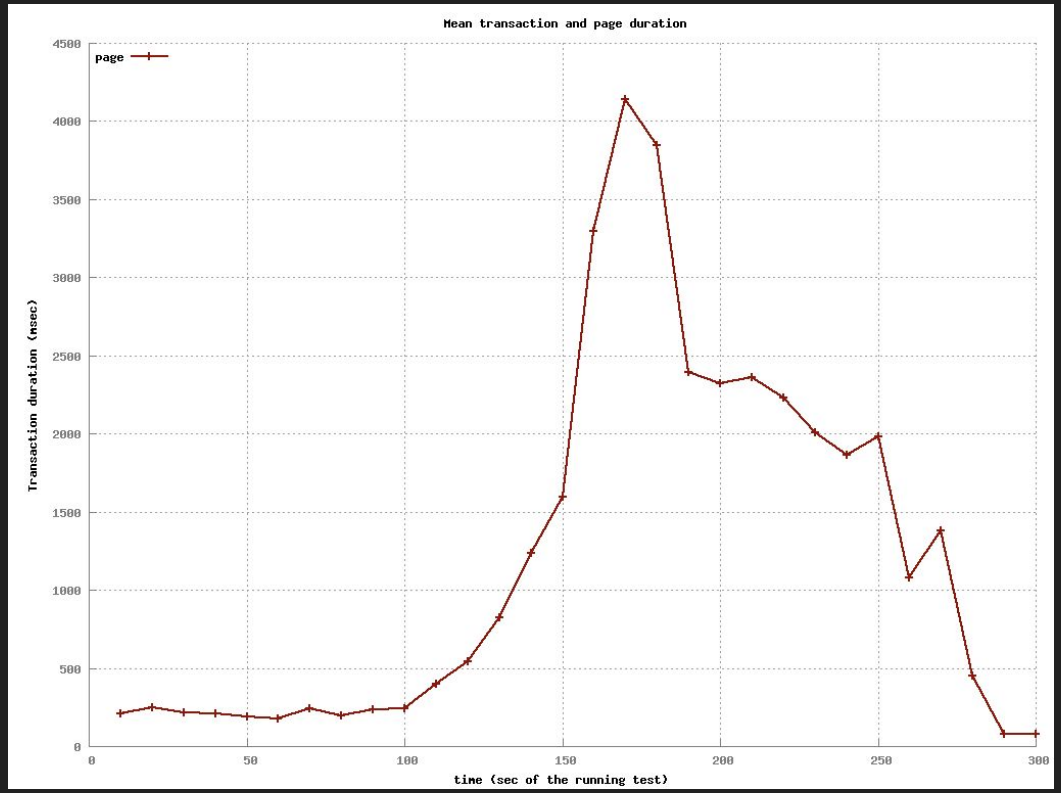
Master branch with memcache


12000
9000
6000
3000
0
Mean Response Time (msec)
60
120
180
240
300
360
Time (sec)
Master
Cache File
Memcache

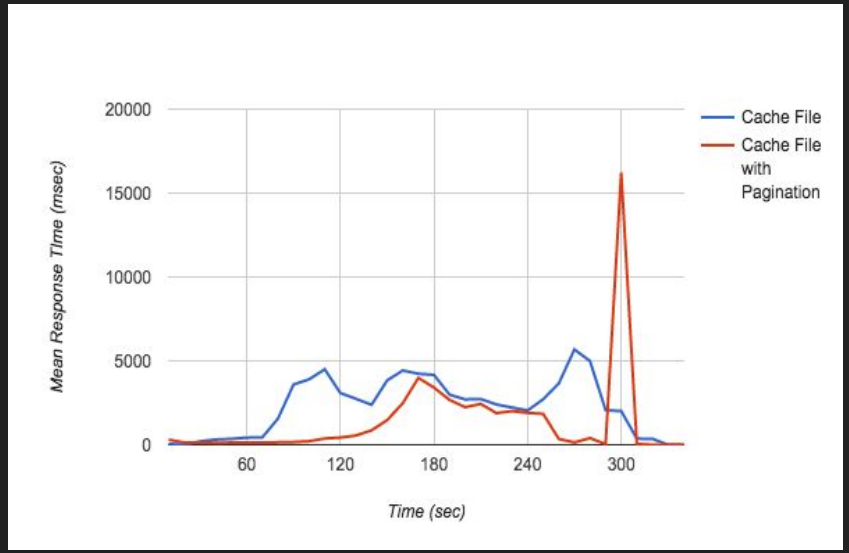
Optimization - Improvement with Pagination in addition to cache file

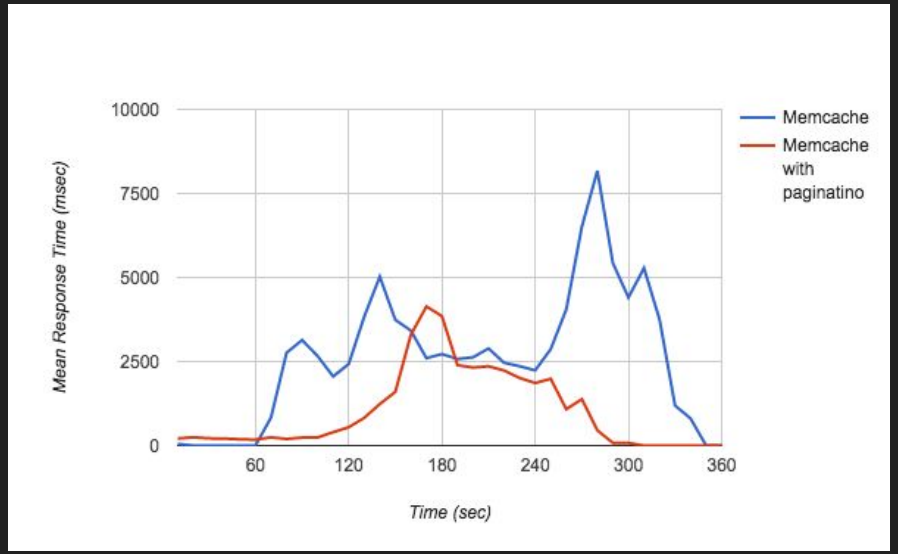


Optimization - Improvement with Pagination in addition to memcache

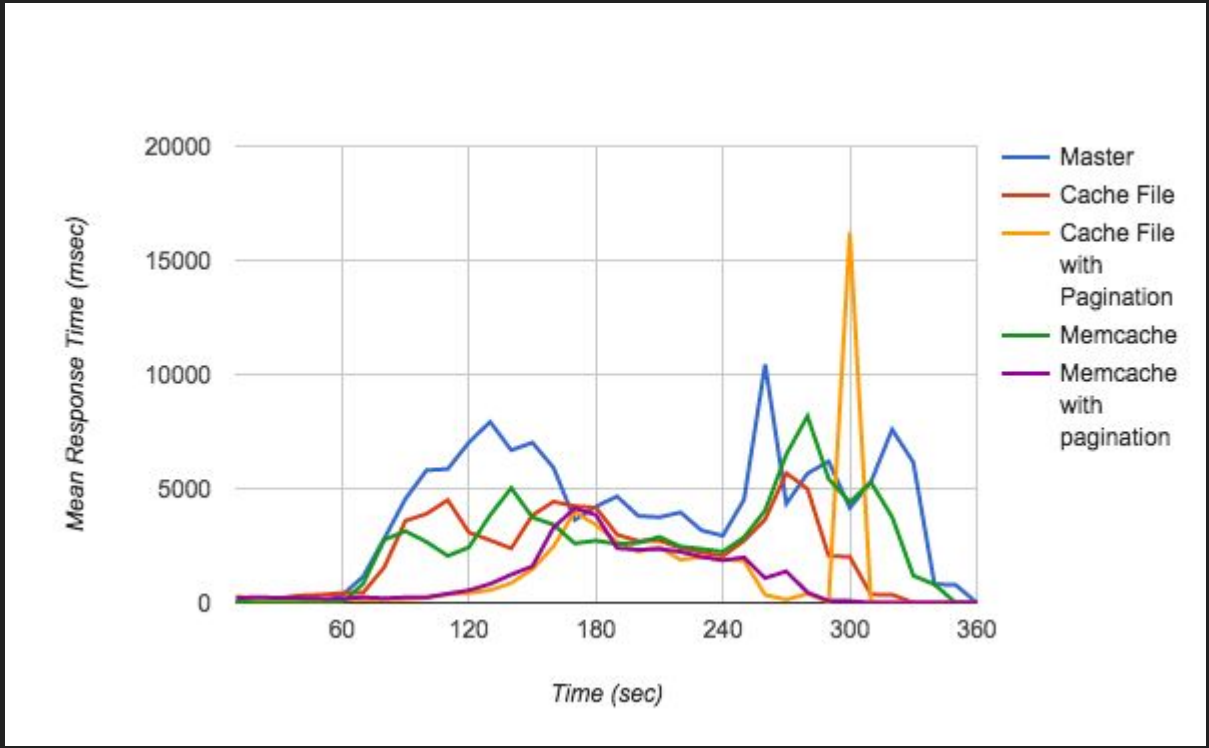


Optimization





Optimization : All experiments



Conclusion

- Applying caching can improve the performance largely.
- Only cache the fragment/action when needed, do not overcache.
- Paginate the result, so the amout of time spent on loading a page would be largely decreased(there will be less time spent on writing to cache for first time loading too)
- memcache does not necessarily perform better than file caching scheme on instances using SSD.
- More optimizations can be done if we look into the process and thread configuration of the instance, and SQL buffer pool.

Thank you very much!

Q & A