

Reddit Engine Technical Report

Tiger Cao

Summary

Successfully implemented reddit engine using actor model in Glean, achieving 6,216 actions/second with 10,000 concurrent users simulated. The system demonstrates good scalability, with independent actor processing each individual user, subreddit, post, comment, and direct message conversations. It also utilizes realistic Zipf-distributed user behavior simulation for joining subreddits.

System Architecture

There are 5 managers for user, subreddit, post, comment, and dm's. Each spawns an actor for every sub-component.

Implementation Challenges & Solutions

Mailbox Overflow:

Problem: With over 20,000 concurrent users, the system sends 100,000+ actions/second but only processes very few per second, causing mailbox congestion.

Solution: increasing delays between messages sent; also later lowered number of users

Actor Initialization Synchronization:

Problem: simulation started before all 10,000 user actors fully initialized, causing timing inaccuracies

Solution: implemented coordinator pattern, only begin simulation after all users ready.

Post Manager Overload:

Post Manager had tens of thousands of live creation; it's also connected to every upvote, downvote, comment, and each requires a dictionary lookup, leading to too much message load; it is involved in 80% of all system messages. It would continue to create posts even after simulation ends.

Solution: decreased user count and rebalanced action distribution.

Zipf Distribution

I implemented Zipf for both subreddit subscription and user activity; specific function implementation are in the reddit.gleam file.

Rank	Subreddit	Subscribers
1	subreddit_1	8,519
2	subreddit_2	7,246
3	subreddit_3	4,191
4	subreddit_4	2,228
5	subreddit_5	1,177

I implemented user activity as a curve as a function of their user id, starting from 1, with a zipf distribution from minimum delay to maximum.

(All stats based on 500 subreddits and 10,000 users)

Action By Category

Action Type	Count	Percentage
DMs Sent	44,444	26.8%
Upvotes	29,640	17.9%
Downvotes	29,738	18.0%
Comments	29,476	17.8%
Subreddit Joins	24,853	15.0%
Create Posts	7,492	4.5%

Performance Results

Users: 10,000

Subreddits: 500

Duration: 20 seconds

Total Actions: 165,643

Action/Second: 8,282.15

This is based on the distribution of action that I picked; again, the bottleneck is the post manager that is involved in every comment, upvote, downvote; to improve this system, it would be good to explore multiple post managers or to change the relationship between different subsystems.

How to run simulation:

```
gleam run reddit.gleam
```