An Exploration of Serverless Architectures for Information Retrieval

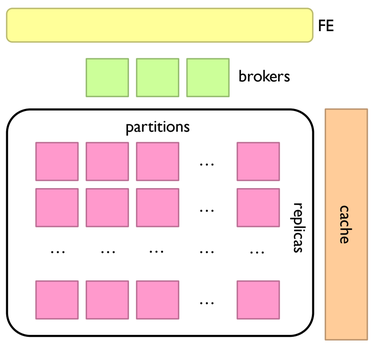
Matt Crane and Jimmy Lin
University of Waterloo

tl; dr – We demonstrate a prototype serverless search engine using Amazon Web Services. Such a design is feasible, but not *yet* practical for interactive querying. The pay-per-query cost model is economically compelling.

Server Architectures

Search engines are built on servers:

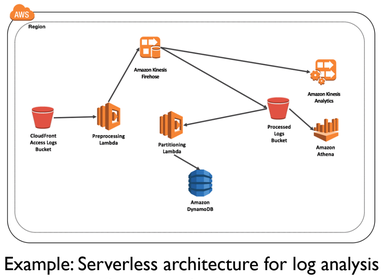
- Services wait for requests (e.g., REST) and respond with results
- Running services requires resource provisioning (even if in the cloud)
- Servers are "always on" – even if not serving any requests
- Elasticity (scaling up and down) is only possible at the server instance level



Serverless Architectures

"Serverless computing": the latest trend in "as a service" cloud computing

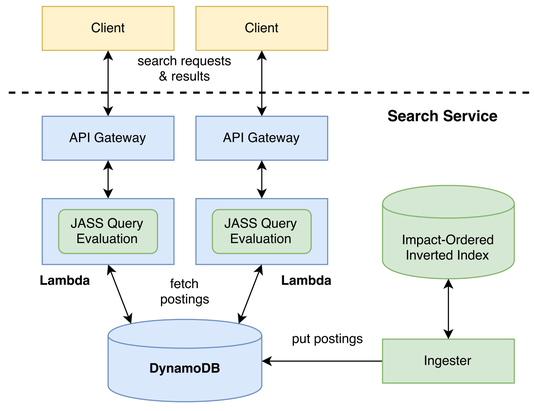
- Applications factored into "state" and "application logic"
- "State" - captured in DB-as-a-service (e.g., Amazon DynamoDB)
- "Application Logic" – executed as stateless functions (e.g., Amazon Lambda)



Example: Serverless architecture for log analysis (Image credit: Amazon)

"Serverless computing" doesn't actually mean you don't need servers…
Just that provisioning, managing, etc. become someone else's problem!

A Serverless Search Engine?



"State" – postings lists

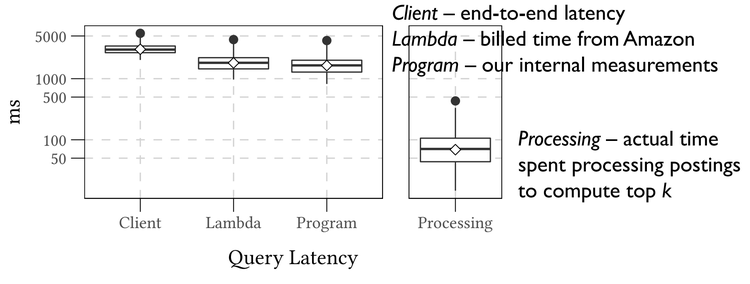
- Impact-ordered index stored in DynamoDB, Amazon's noSQL key-value store
- Schema: index term = key, postings = value
- Hack to circumvent size restrictions in DynamoDB

"Application Logic" – query evaluation

- JASS query evaluation algorithm for impact-ordered indexes
- API gateway calls implementation: looks up postings, processes them to compute top *k*.

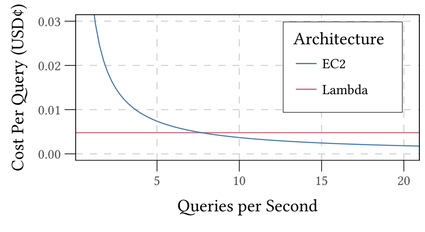
Experiments

Prototype on Gov2 collection (25 million docs), topics 701-850



Client – end-to-end latency
Lambda – billed time from Amazon
Program – our internal measurements

Processing – actual time spent processing postings to compute top *k*



- Feasible, but currently not practical for interactive querying (~3s query latency)
- Everything other than "Processing" is serverless overhead: will get better!

- Cost per query: USD $0.000047951
- Dedicated EC2 instance – cost per query varies by load
- Cross-over point at 7.7 QPS