HaaS Architecture

gurjeet
@supabase.io
@singh.im

Ideal Features of a PGaaS

- Postgres Compatibility
- Synchronized Releases with Community
- Point-in-Time Recovery
- Automatic Daily Backups
- Manual/Named Backups
- Database Instance Branches/Clones
- Encryption-at-Rest

Ideal Features (contd.)

- Online Storage Auto-Scaling
- Online Compute (and RAM) Auto-Scaling
- Online Version Upgrades
- Scale-to-Zero
- Ability to Throttle IOPS
- Offline Support

HaaS

- An attempt at implementing the ideal features
- Haathi-as-a-Service
  - Haathi: 'Elephant', in Hindi

Obligatory Denial

ElepHaaS: Not What You Think

HaaS != Haas

Shaun M. Thomas

bonesmoses.org/presentations/PostgresOpen2016_ElepHaaS/

Elephant Herd as a Service
Postgres Open 2016

Sorry to all the Haas groupies!

Research

- Investigated various open-source projects
- Read about various products (commercial and open-source)
- Compared their architectures
- Brainstormed ideas and architectures
- Proposed architecture chosen after considering many others

Research (Contd.)

- Postgres-XC (Postgres-XL, Tbase)
- Postgres on OpenEBS (userland OpenZFS)
- Amazon Aurora
- CitusDB, bought by Microsoft
- PolarDB, by Alibaba
- Oracle Exadata, Oracle RAC
- Google Spanner
- TimescaleDB
- Modify an open-source iSCSI driver to do ***our*** thing
- ... others I may be forgetting

Warning

- This is a Moonshot Project

  Inspired after seeing Google Spanner demo

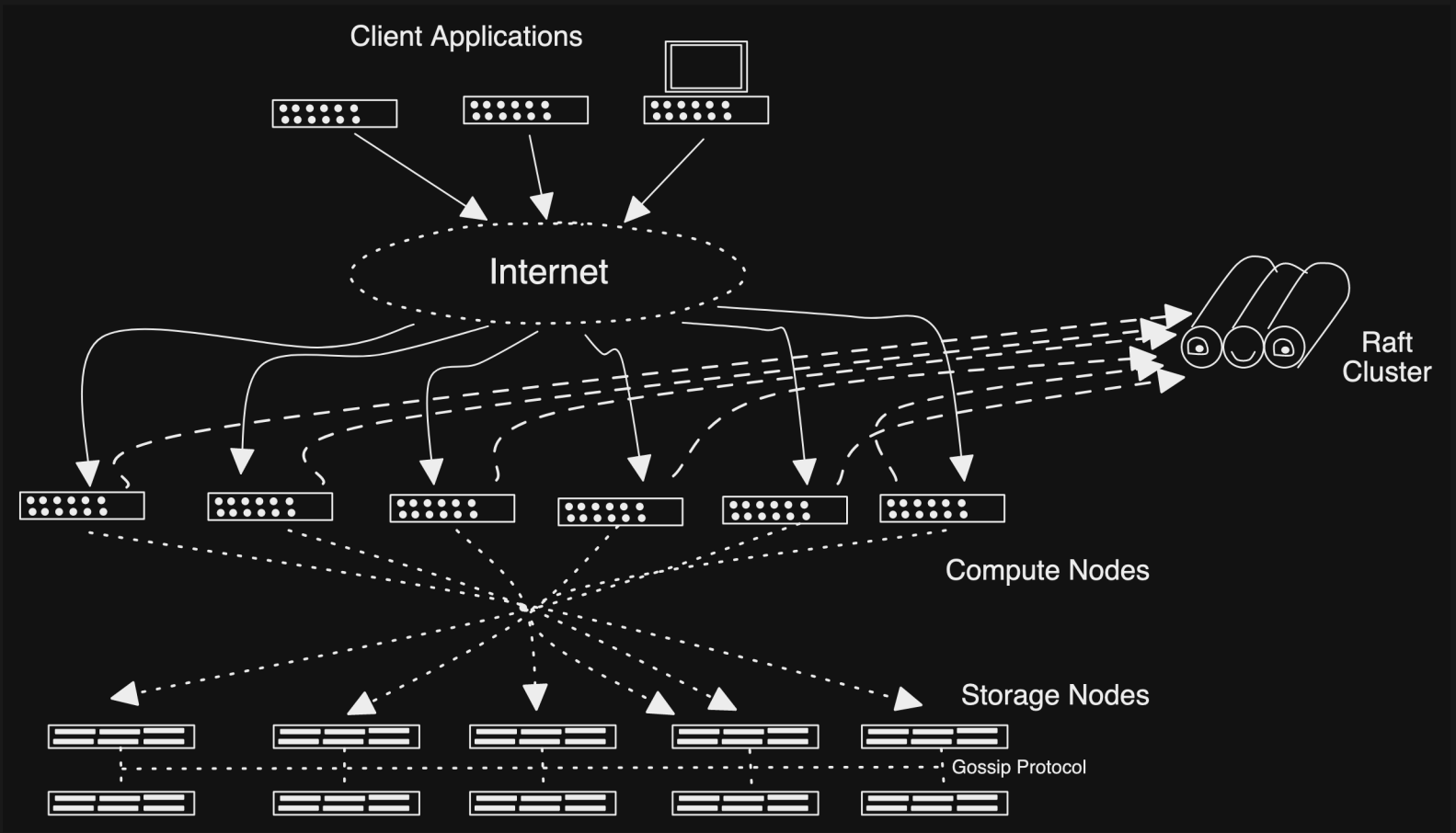
  Quote: *... if we're aiming for the moon, I think we should aim to be more like Google Spanner, rather than try to emulate Amazon Aurora.*

  from the email to @copple, @ant, and @inian.
- The implementation details are hand-wavy as of now

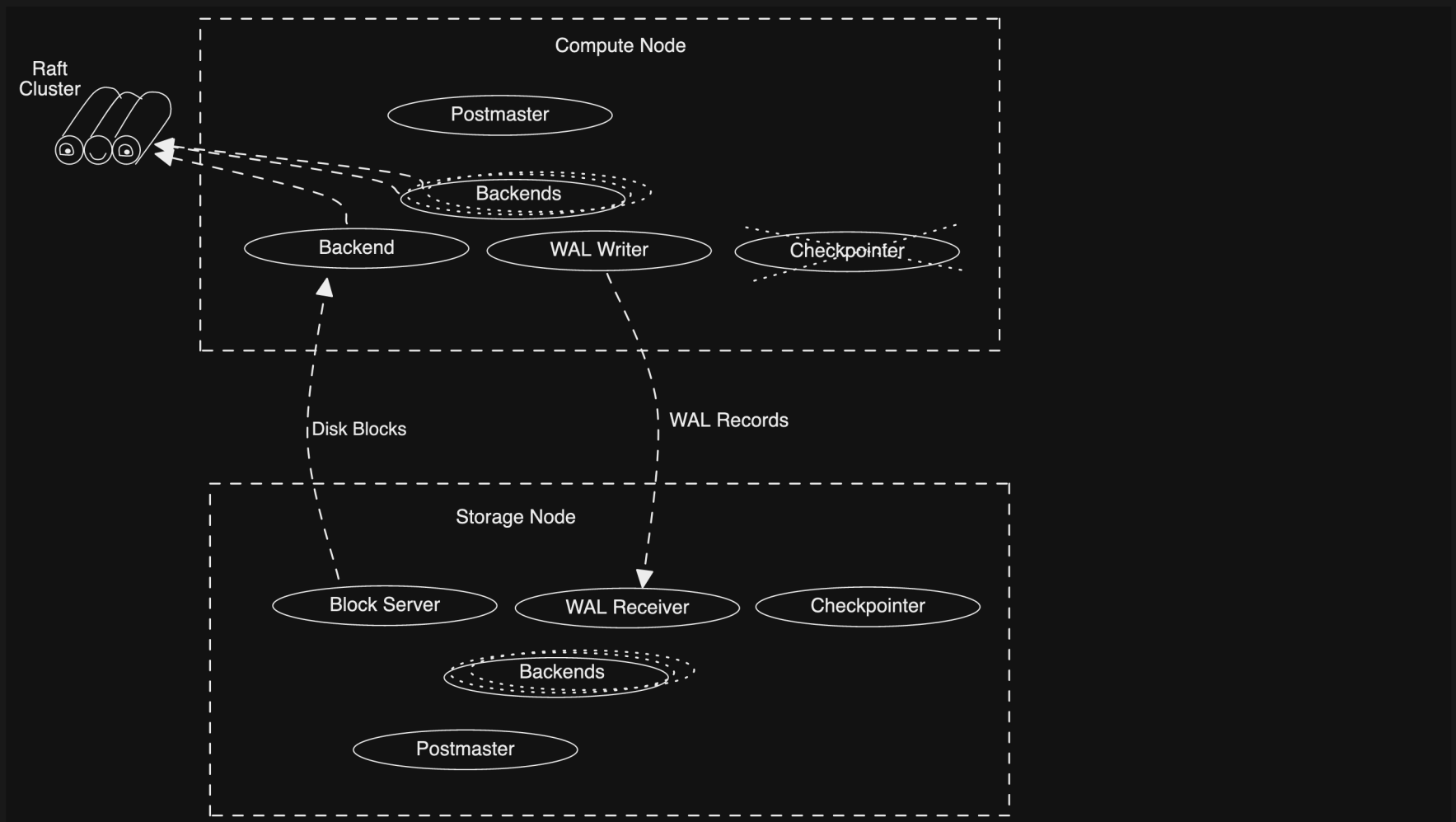
HaaS Architecture in a *Nutshell*

- One or more Postgres primary DB Instances
  - Compute Nodes
- One or more Postgres standby DB Instances
  - Storage Nodes
- Compute Nodes send Write-Ahead Log (WAL) to Storage Nodes
- Compute Nodes do ***not*** write to local filesystem
- Compute Nodes read blocks from Storage Nodes

HaaS Arch. in a Thousand Words



HaaS Arch. Drilldown



HaaS Architecture - More Details

- A Raft protocol based Lock Manager
  - External Lock Manager (ELM)
- Compute Nodes use ELM to take shared or exclusive locks on blocks
- No need to consult ELM when there's just one Compute Node

More Details (Contd.)

- The Storage Nodes
  - use a Gossip protocol to communicate w/ each other
  - take ownership of 1GB-sized segments
  - forward relevant WAL records to peers
  - ensure Durability (from ACID) guarantee

A Note About Performance

- Correctness above speed/performance
- Performance is an invisible feature
- There are only 2 kinds of databases
  - the ones that have great features, and
  - the ones that are used by the customers.

A Note About Perf. (Contd.)

- The single-node MVP version's performance
  - should be comparable to Postgres
    - on same or similar hardware, or
    - on different hardware, but at the same price-point

Milestone 1

- Modify Postgres to:
  - make it use a standby for all data/files
  - serve read requests made by compute node

Milestone 1 - Evaluation Criteria

- Successfully run pgbench
  - in read-write mode
  - with single client connection
  - for at least 60 seconds
  - that completes at least 1 transaction
  - with 0 errors

Future

(waa..y into the future)

- Predicate/Aggregate push-down to Storage Nodes
  - process WHERE clause closer to storage
  - process GROUP BY clause closer to storage
- Leverage RDMA
  - RDMA in Cloud becoming a reality
  - Amazon EC2 Elastic Fabric Adapter (EFA)

Questions/Comments

gurjeet
@supabase.io
@singh.im