Cloud Cascade

cloudcascade.io

Shardool Pathak - Founder
contact@cloudcascade.io
https://www.linkedin.com/in/shardool-p-86a066125/

How to
Seamlessly Automate and Abstract away HPC Infrastructure

Problem Context

Current HPC infra and cluster setup involves…

Steep Learning Curve: Weeks of training needed to master cloud-specific HPC deployment(AWS, GCP)

Problem Context

Current HPC infra and cluster setup involves...

Steep Learning Curve: Weeks of training needed to master cloud-specific HPC deployment(AWS, GCP)

Custom Effort Required: Each cloud provider demands a unique setup and configuration

Problem Context

Current HPC infra and cluster setup involves...

Steep Learning Curve: Weeks of training needed to master cloud-specific HPC deployment(AWS, GCP)

Custom Effort Required: Each cloud provider demands a unique setup and configuration

Non-Transferable Skills: Knowledge does not easily transfer between different cloud platforms

Problem Context

Current HPC infra and cluster setup involves...

Steep Learning Curve: Weeks of training needed to master cloud-specific HPC deployment(AWS, GCP)

Custom Effort Required: Each cloud provider demands a unique setup and configuration

Non-Transferable Skills: Knowledge does not easily transfer between different cloud platforms

High Costs: User are often surprised with high hourly costs for clusters post-deployment

Central Idea

Effortless HPC cluster deployment for scientific computing, research, ML/AI model training etc.

Simplicity: intuitive, natural language interactions

Central Idea

Effortless HPC cluster deployment for scientific computing, research, ML/AI model training etc.

Simplicity: intuitive, natural language interactions

Freedom: Work seamlessly across any cloud (AWS, Google Cloud, Azure, Oracle etc.)

Central Idea

Effortless HPC cluster deployment for scientific computing, research, ML/AI model training etc.

Simplicity: intuitive, natural language interactions

Freedom: Work seamlessly across any cloud (AWS, Google Cloud, Azure, Oracle etc.)

Speed: let AI take care of infra complexity, zero learning curve

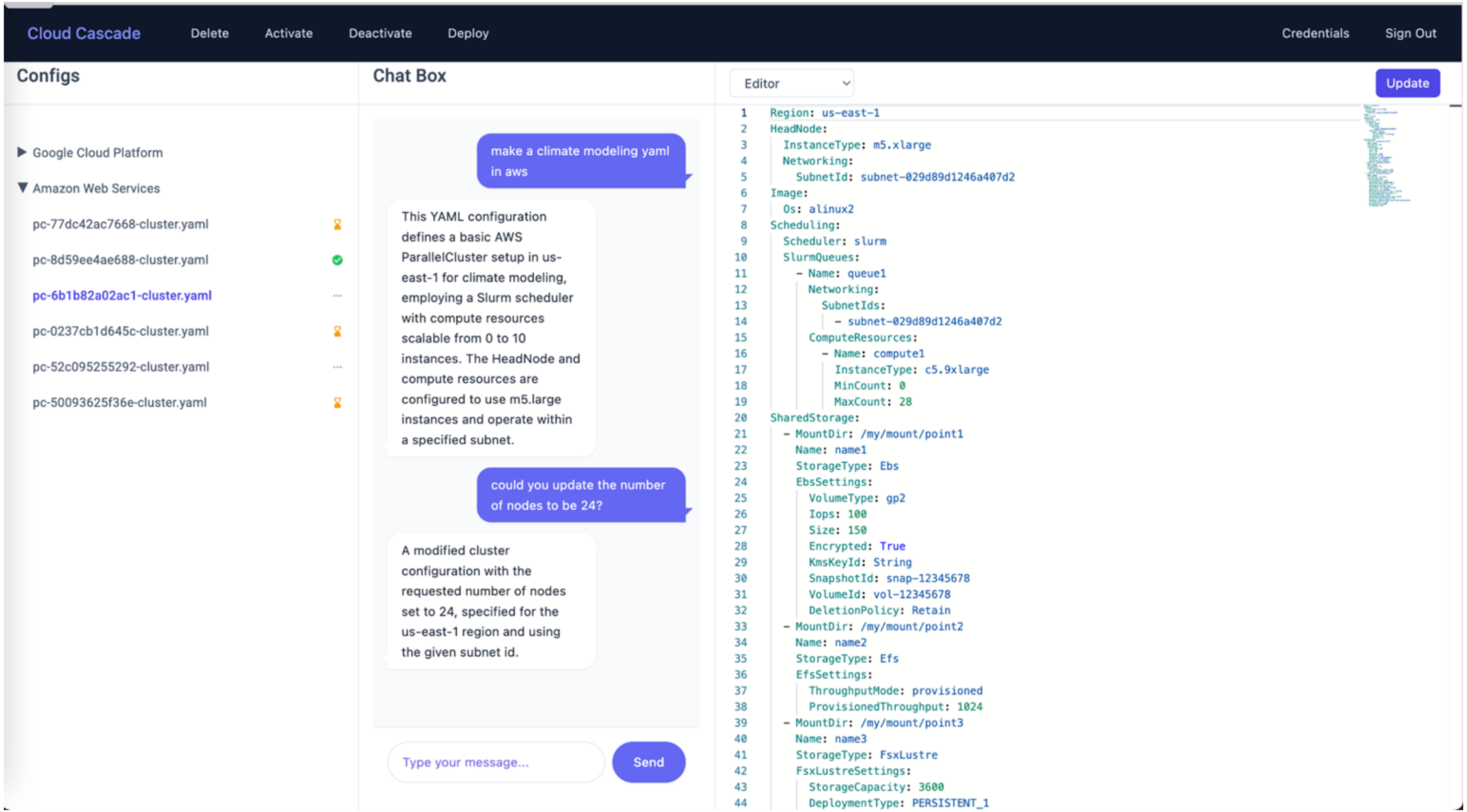
Features

- Automated Config Generation w/ Validation

- Pre-Deployment Cost Breakdown

- Seamless Cluster Management via UI

Automated Config Generation w/ Validation



Configs

▶ Google Cloud Platform

▼ Amazon Web Services

- pc-77dc42ac7668-cluster.yaml ⌛
- pc-8d59ee4ae688-cluster.yaml ✓
- pc-6b1b82a02ac1-cluster.yaml ⋯
- pc-0237cb1d645c-cluster.yaml ⌛
- pc-52c095255292-cluster.yaml ⋯
- pc-50093625f36e-cluster.yaml ⌛

Chat Box

make a climate modeling yaml in aws

This YAML configuration defines a basic AWS ParallelCluster setup in us-east-1 for climate modeling, employing a Slurm scheduler with compute resources scalable from 0 to 10 instances. The HeadNode and compute resources are configured to use m5.large instances and operate within a specified subnet.

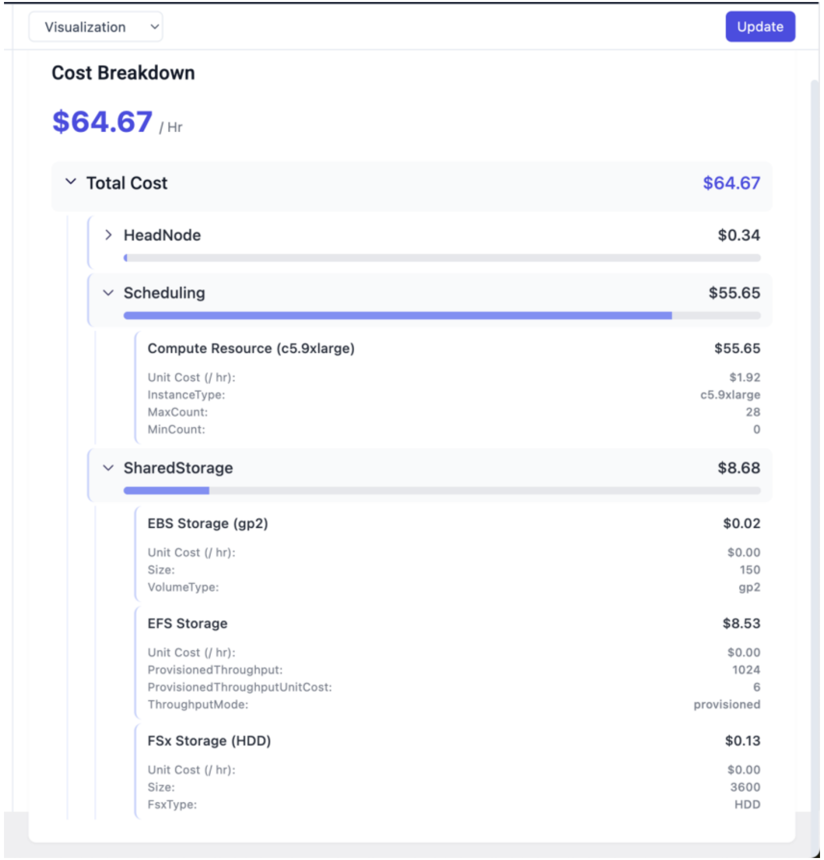
could you update the number of nodes to be 24?

A modified cluster configuration with the requested number of nodes set to 24, specified for the us-east-1 region and using the given subnet id.

Type your message...

```
Region: us-east-1
HeadNode:
  InstanceType: m5.xlarge
  Networking:
    SubnetId: subnet-029d89d1246a407d2
Image:
  Os: alinux2
Scheduling:
  Scheduler: slurm
  SlurmQueues:
    - Name: queue1
      Networking:
        SubnetIds:
          - subnet-029d89d1246a407d2
      ComputeResources:
        - Name: compute1
          InstanceType: c5.9xlarge
          MinCount: 0
          MaxCount: 28
SharedStorage:
  - MountDir: /my/mount/point1
    Name: name1
    StorageType: Ebs
    EbsSettings:
      VolumeType: gp2
      Iops: 100
      Size: 150
      Encrypted: True
      KmsKeyId: String
      SnapshotId: snap-12345678
      VolumeId: vol-12345678
      DeletionPolicy: Retain
  - MountDir: /my/mount/point2
    Name: name2
    StorageType: Efs
    EfsSettings:
      ThroughputMode: provisioned
      ProvisionedThroughput: 1024
  - MountDir: /my/mount/point3
    Name: name3
    StorageType: FsxLustre
    FsxLustreSettings:
      StorageCapacity: 3600
      DeploymentType: PERSISTENT_1
```

Pre-Deployment Cost Breakdown



Cost Breakdown

$64.67 / Hr

Total Cost $64.67

HeadNode $0.34

Scheduling $55.65

Compute Resource (c5.9xlarge) $55.65

Unit Cost (/ hr): $1.92
InstanceType: c5.9xlarge
MaxCount: 28
MinCount: 0

SharedStorage $8.68

EBS Storage (gp2) $0.02

Unit Cost (/ hr): $0.00
Size: 150
VolumeType: gp2

EFS Storage $8.53

Unit Cost (/ hr): $0.00
ProvisionedThroughput: 1024
ProvisionedThroughputUnitCost: 6
ThroughputMode: provisioned

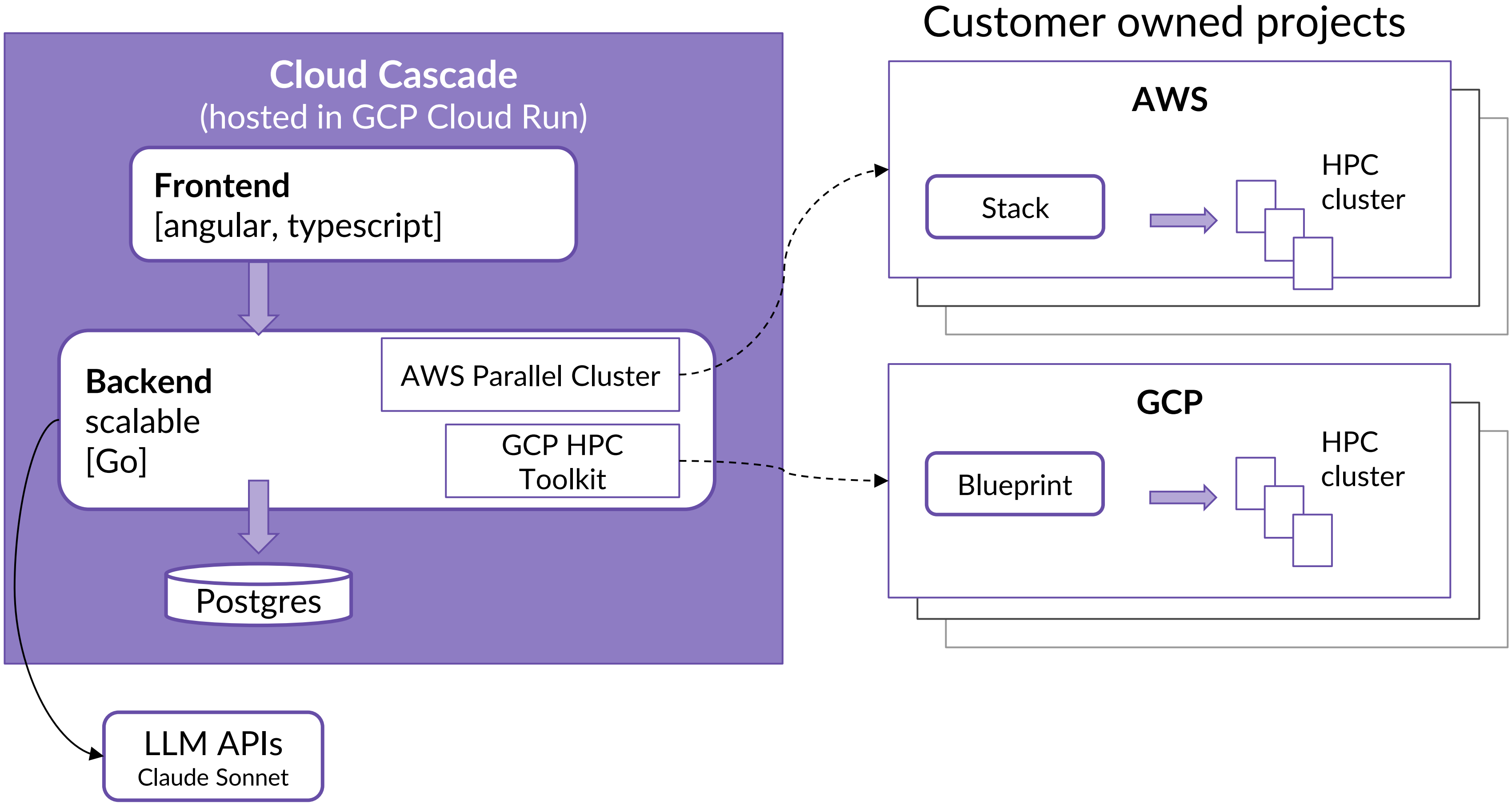
FSx Storage (HDD) $0.13

Unit Cost (/ hr): $0.00
Size: 3600
FsxType: HDD

Demo

https://youtu.be/G9WWXuJiGeE?si=IW9eYZhfKwY9BNQd

Architecture



Next Features

More Cloud Support (Azure, NVidia, Oracle)

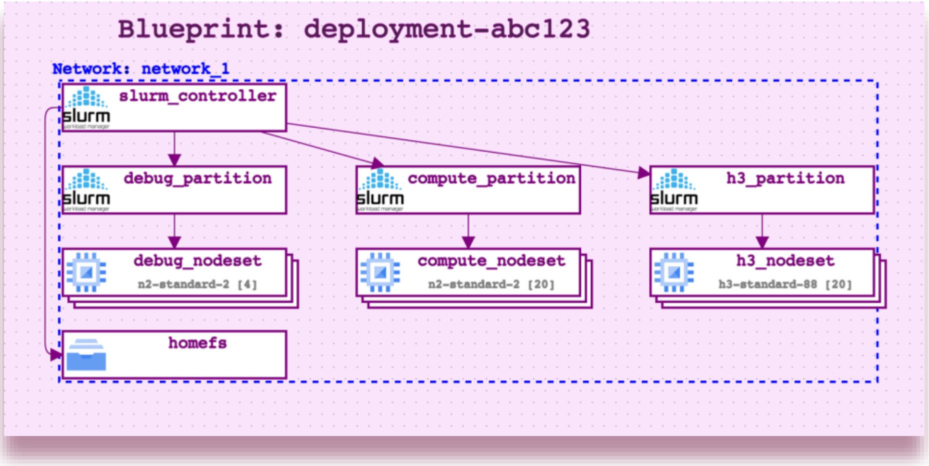
Cloud to Cloud Conversion (AI agent)

Machine / core equivalence

Price equivalence

Cluster Diagrams & Visualization

Cluster Diagrams & Visualization - GCP Blueprint



Questions?