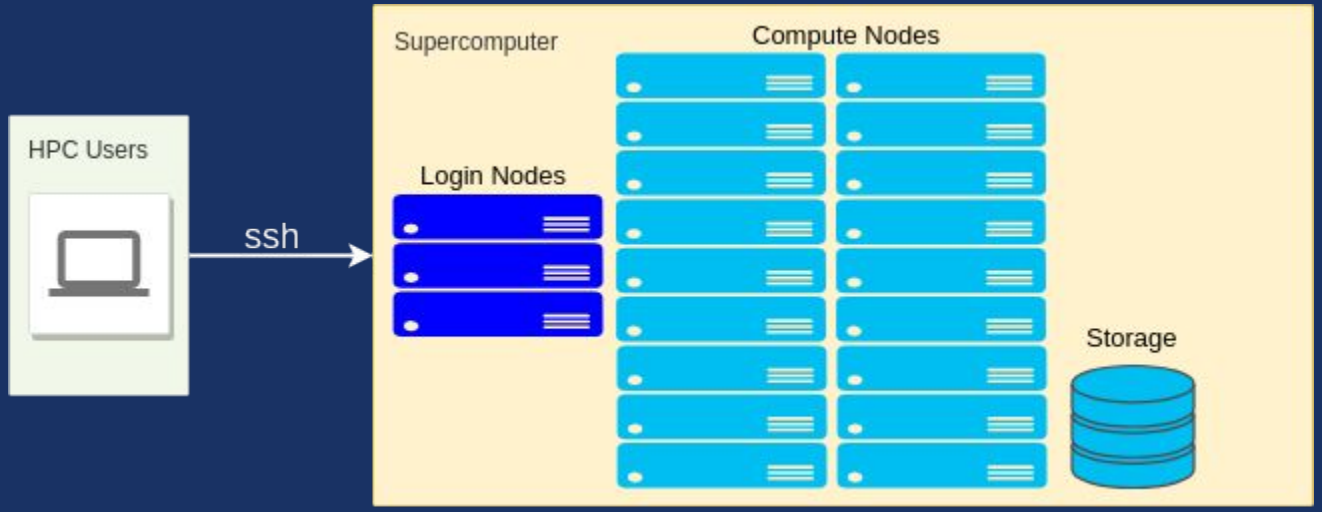
Bare-Metal Style HPC Clusters On Google Cloud Platform

Dr. Joseph Schoonover

CEO, Cloud-HPC Systems Architect
Fluid Numerics, LLC
Boulder, CO

joe@fluidnumerics.com

FluidNumerics



This talk will show how we can replicate this type of environment on the cloud

FluidNumerics

HPC Components

Compute

Network

Storage

FluidNumerics

Deployment Manager

https://cloud.google.com/deployment-manager/

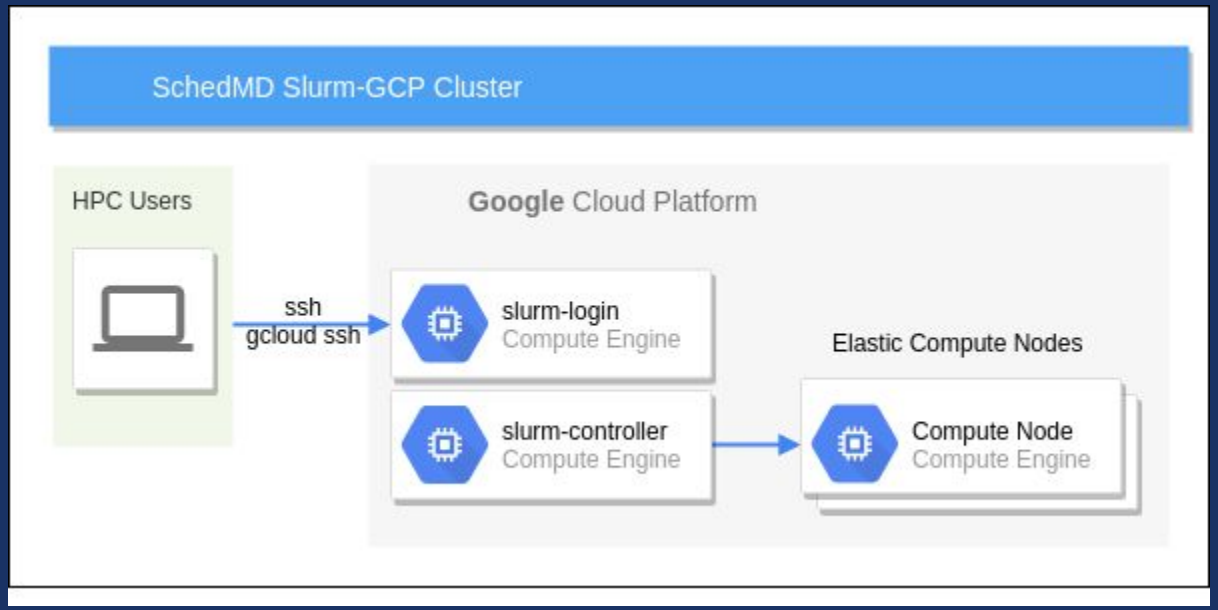
Infrastructure as code
→ Python or Jinja templates
and YAML dictionaries

```
Resources:
- type: compute.v1.instance
  name: quickstart-deployment-vm
  Properties:
    zone: us-central1-f
    machineType: <MACHINE TYPE>
    Disks:
   - deviceName: boot
      type: PERSISTENT
      boot: true
      autoDelete: true
      initializeParams:
        sourceImage: <OS IMAGE>
   networkInterfaces:
    - network: <NETWORK>
```

FluidNumerics

SchedMD : Slurm-GCP

https://github.com/schedmd/slurm-gcp



FluidNumerics

SchedMD : Slurm-GCP

How to use : Customize YAML file

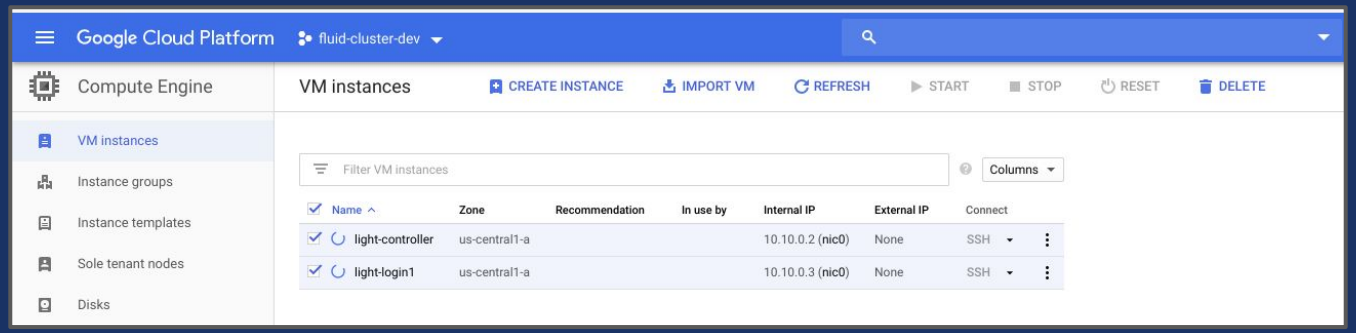
```
#slurm-cluster.yaml

imports:
- path: slurm.jinja
resources:
- name: slurm-cluster
  type: slurm.jinja
  properties:
    cluster_name            : g1
    static_node_count       : 2
    max_node_count          : 10
    zone                    : us-central1-b
    region                  : us-central1
    cidr                    : 10.10.0.0/16
    controller_machine_type : n1-standard-2
    compute_machine_type    : n1-standard-2
    login_machine_type      : n1-standard-2
```

FluidNumerics

SchedMD : Slurm-GCP

How to use : Deploy

```
gcloud deployment-manager deployment create slurm-cluster --project=<project id> --config=slurm-cluster.yaml
```

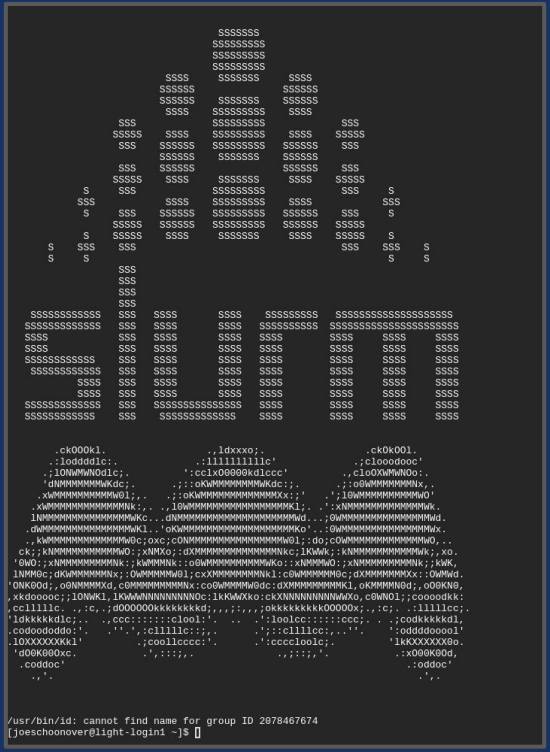


FluidNumerics

SchedMD : Slurm-GCP

How to use : Login

```
ssh <user>@<ip-address>
```



FluidNumerics

Customizations

- Stackdriver monitoring integrations
- Multiple, user-locked, micro-login nodes
- Additional NFS storage and Lustre Storage
- Multiple partitions for multiple application/developer support
- Federation - burst expand on prem resources
- Python/Jupyter notebook integrations

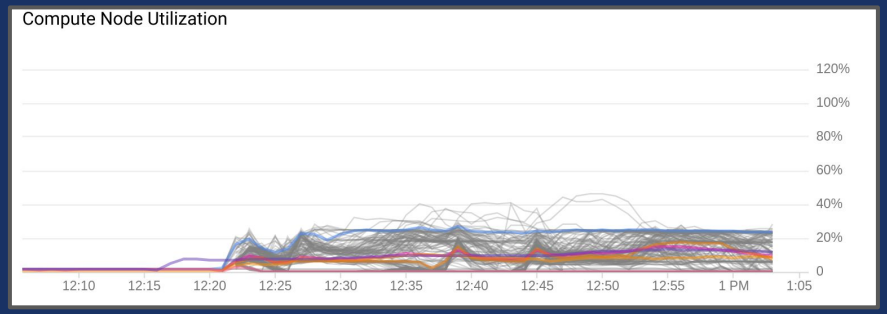
FluidNumerics

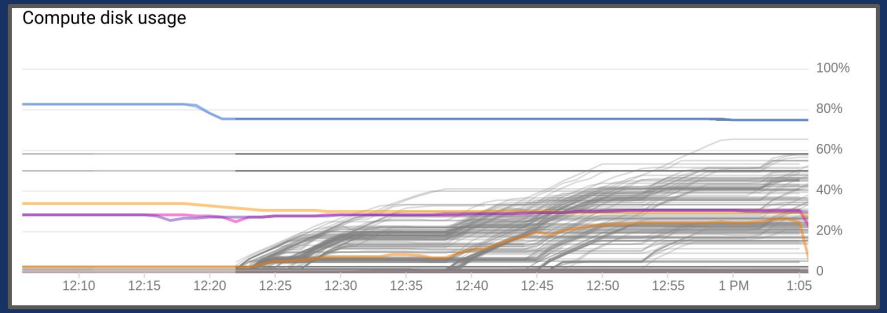
Stackdriver

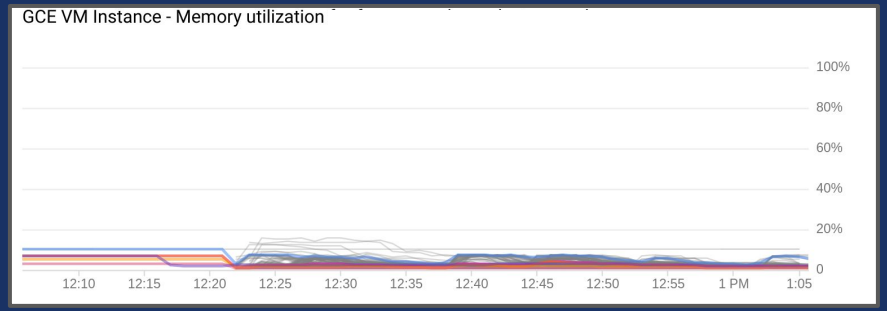
Built in system monitoring and alerting platform

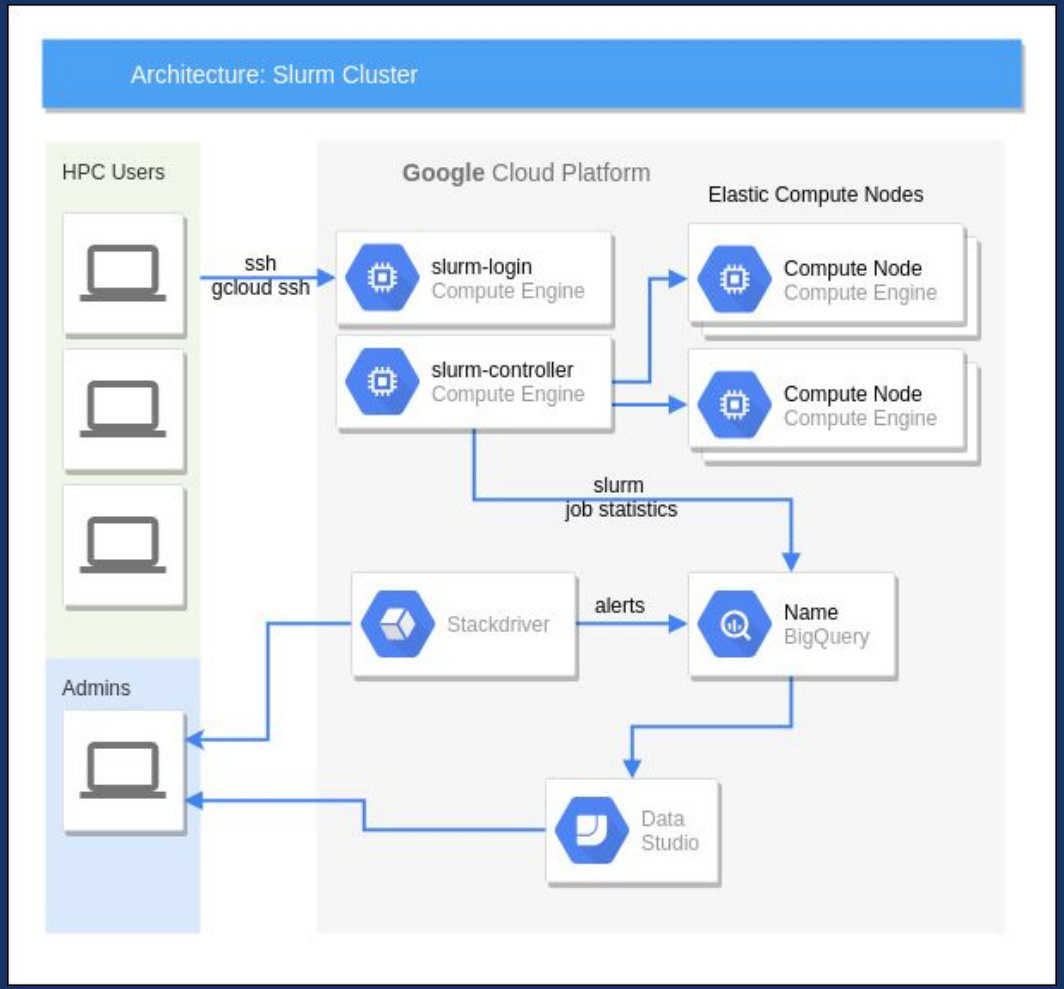
Custom metrics can be integrated (e.g. GPU utilization)

This data can be used to resize compute nodes to maximize resource utilization and reduce costs









Compute Resources

Slurm Cluster

github.com/schedmd/slurm-gcp

Customizations

+multi-partition +multi-zone

+environment modules

+github.com/spack/spack

FluidNumerics

Slurm + Stackdriver + BigQuery + DataStudio

Job Performance by User

User: joe, jodie

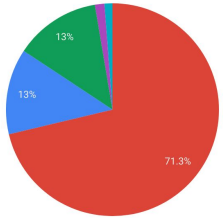
(2)

Individuals and interactions over processes and tools.

Select date range

Build projects around motivated individuals.
Give them the environment and support they need,
and trust them to get the job done.

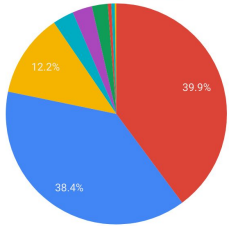
The most efficient and effective method of
conveying information to and within a development
team is face-to-face conversation.



- COMPLETED
- CANCELLED by 1045
- RUNNING
- NODE_FAIL
- FAILED

Condition Name	Count
[Moderate] Compute Node Under-utilization Upper Bound	2
[Moderate] Compute Node Under-utilization Lower Bound	2
[Severe] Under-utilized CPU Memory Upper Bound	33
[Serious] Under-utilized CPU Memory	52
[Serious] Under-utilized CPU Memory Lower Bound	52
[Severe] Compute Node Under-utilization	53
[Serious] Compute Node Under-utilization Upper Bound	55
[Serious] Compute Node Under-utilization Lower Bound	56
[Moderate] Compute Node Under-utilization Lower Bound	291
[Moderate] Compute Node Under-utilization Upper Bound	291

1 - 10 / 18



- [Serious] Under-utilized CPU Memory
- [Moderate] Under-utilized CPU Memory
- [Severe] Compute Node Under-Utilization
- [Serious] Compute Node Under-utilization
- [Moderate] Compute Node Under-utilization
- [Severe] Under-utilized CPU Memory
- [Serious] Under-utilized CPU Memory
- [Serious] Compute Node Under-utilization
- [Severe] Compute Node Under-Utilization

Policy Name	User	Partition	Job Name
[Severe] Under-utilized CPU Memory	joe	11	821J9FP01161[FP3]
[Severe] Under-utilized CPU Memory	joe	11	821J9FP0912[FP3]
[Severe] Under-utilized CPU Memory	jodie	00	821J9FP01308[FP3]
[Severe] Under-utilized CPU Memory	jodie	00	821J9FP01151[FP3]
[Severe] Under-utilized CPU Memory	joe	11	821J9FP01259[FP3]

1 - 5 / 776

FluidNumerics

Use Case

GPU Platform Comparisons

Partitions

highmem-64 (64 CPU + 416 GB RAM) + 8 Nvidia® Tesla® V100 GPUs

highmem-32 (32 CPU + 208 GB RAM) + 4 Nvidia® Tesla® P100 GPUs

highmem-16 (16 CPU + 104 GB RAM) + 4 Nvidia® Tesla® P4 GPUs

highmem-16 (16 CPU + 104 GB RAM) + 4 Nvidia® Tesla® K80 GPUs

highmem-8 (8 CPU + 52 GB RAM) + 1 Nvidia® Tesla® K80 GPU

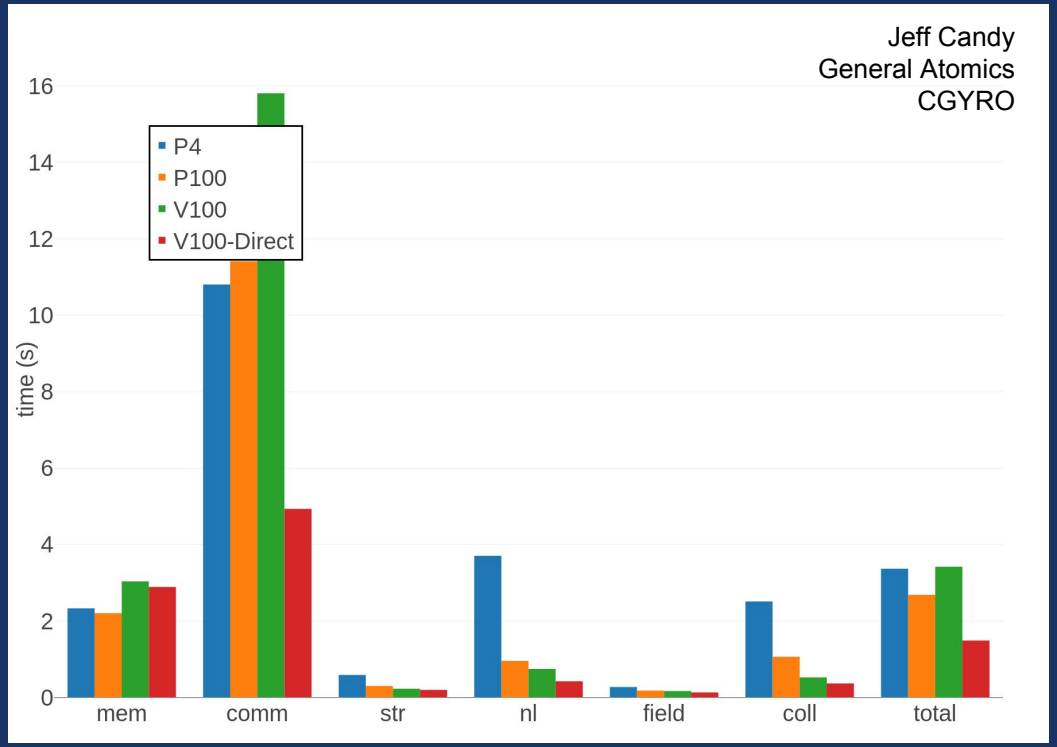
highmem-8 (8 CPU + 52 GB RAM) + 1 Nvidia® Tesla® P100 GPU

standard-32 (32 CPU + 120 GB RAM)

FluidNumerics

Use Case

GPU Platform Comparisons

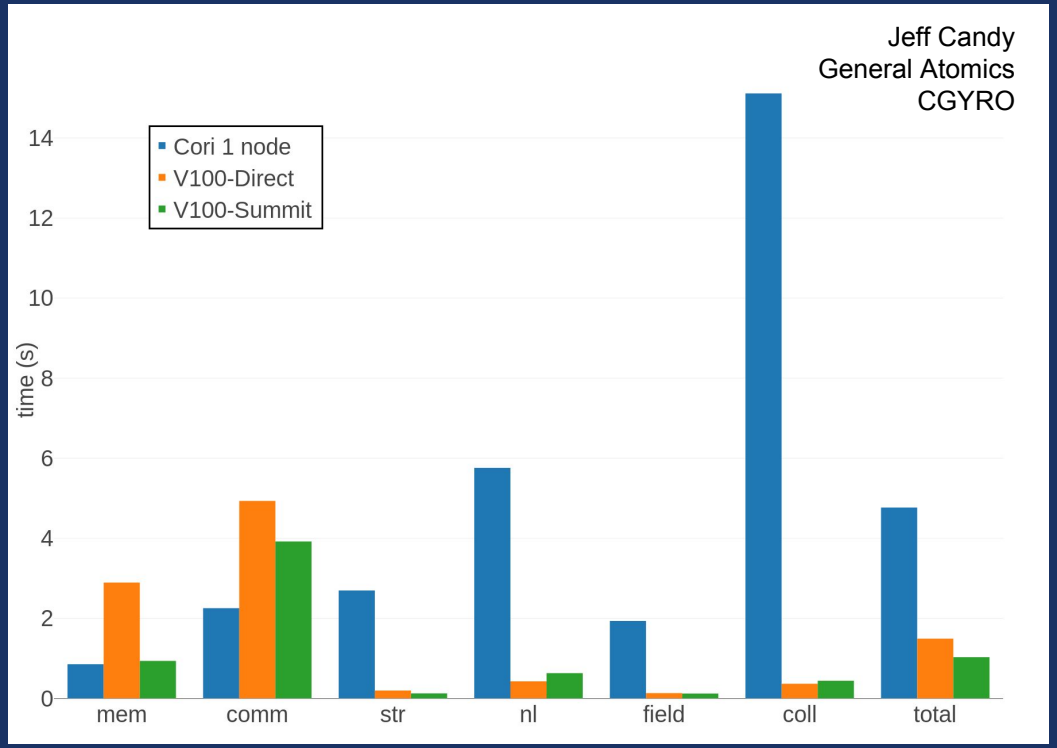


- Highmem-32 + 4xV100 GPU
- Highmem-32 + 4xP100 GPU
- Highmem-32 + 4xV100 GPU
- Highmem-32 + 4xV100 GPU + GPU-Direct MPI

FluidNumerics

Use Case

Cori, Summit, and GCP Comparisons

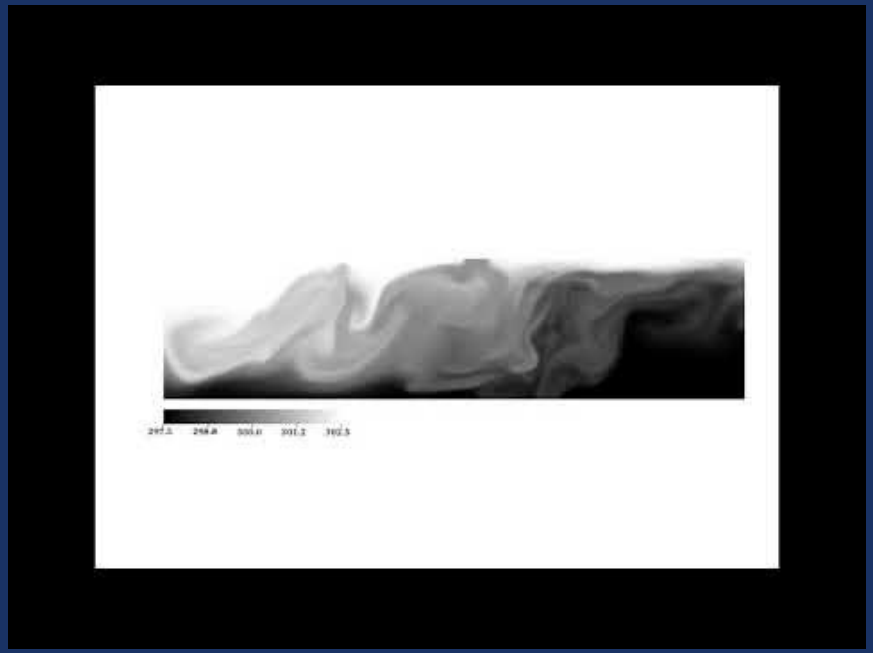


- Highmem-32 + 4xV100 GPU
- Highmem-32 + 4xP100 GPU
- Highmem-32 + 4xV100 GPU
- Highmem-32 + 4xV100 GPU + GPU-Direct MPI

FluidNumerics

Use Case

Cloud Cost Analysis

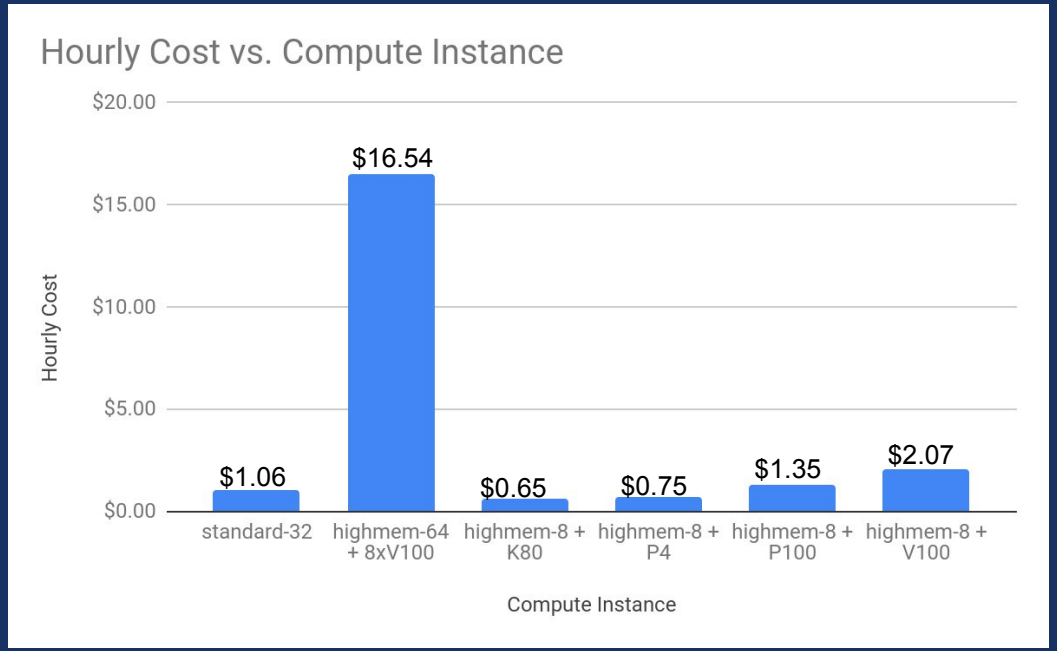


- Standard-32
- Highmem-8 + 1xP4 GPU
- Highmem-8 + 1xP100 GPU
- Highmem-8 + 1xV100 GPU
- Highmem-64 + 8xV100 GPU

FluidNumerics

Use Case

Cloud Cost Analysis

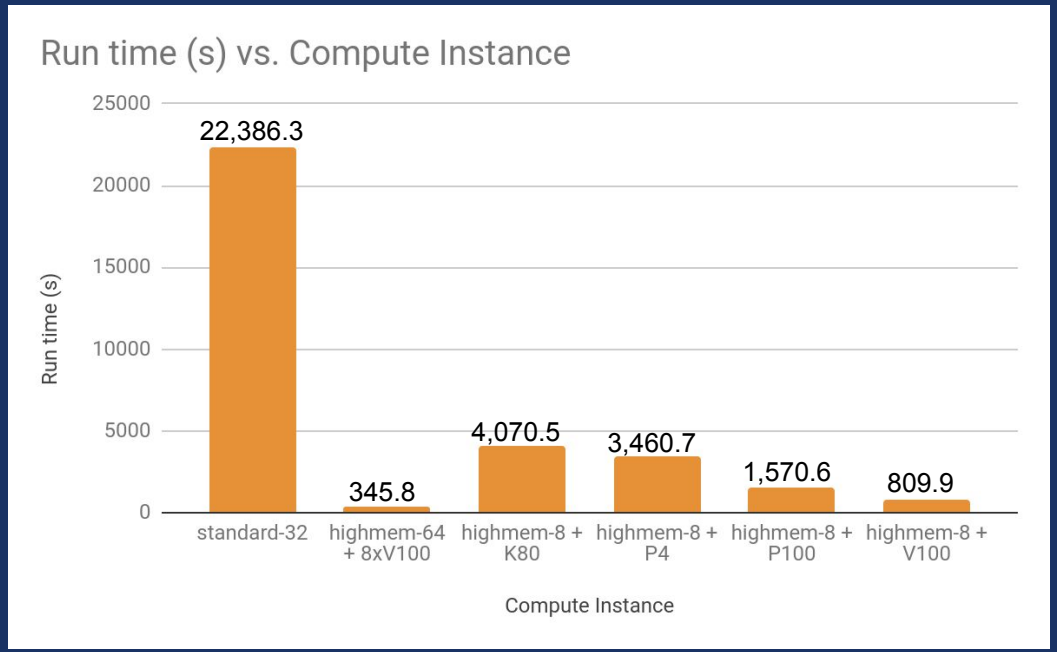


- Standard-32
- Highmem-8 + 1xP4 GPU
- Highmem-8 + 1xP100 GPU
- Highmem-8 + 1xV100 GPU
- Highmem-64 + 8xV100 GPU

FluidNumerics

Use Case

Cloud Cost Analysis

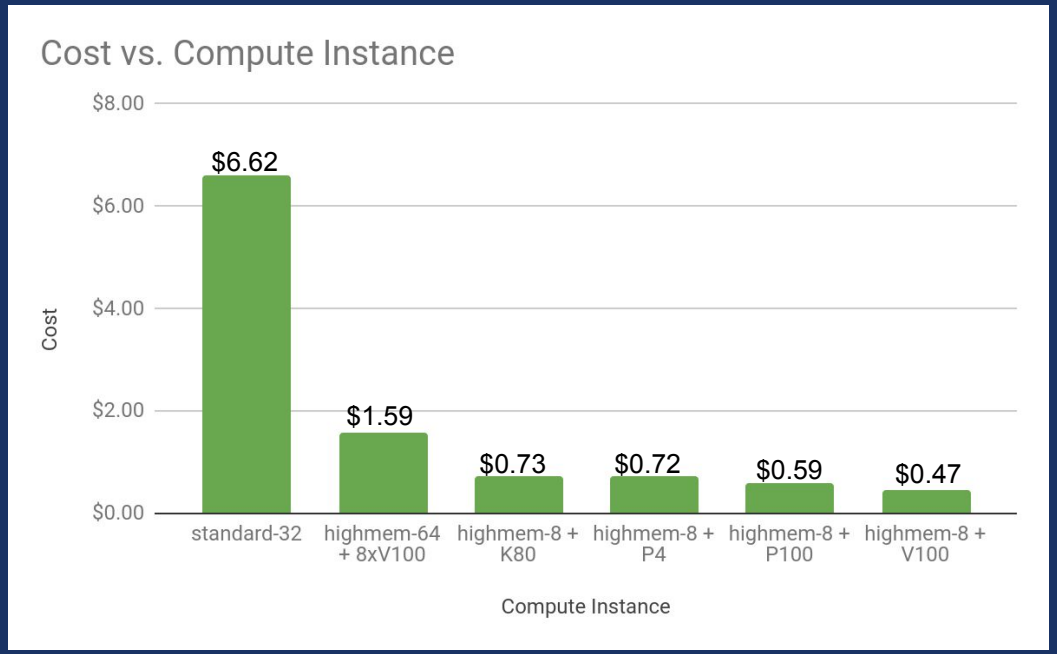


- Standard-32
- Highmem-8 + 1xP4 GPU
- Highmem-8 + 1xP100 GPU
- Highmem-8 + 1xV100 GPU
- Highmem-64 + 8xV100 GPU

FluidNumerics

Use Case

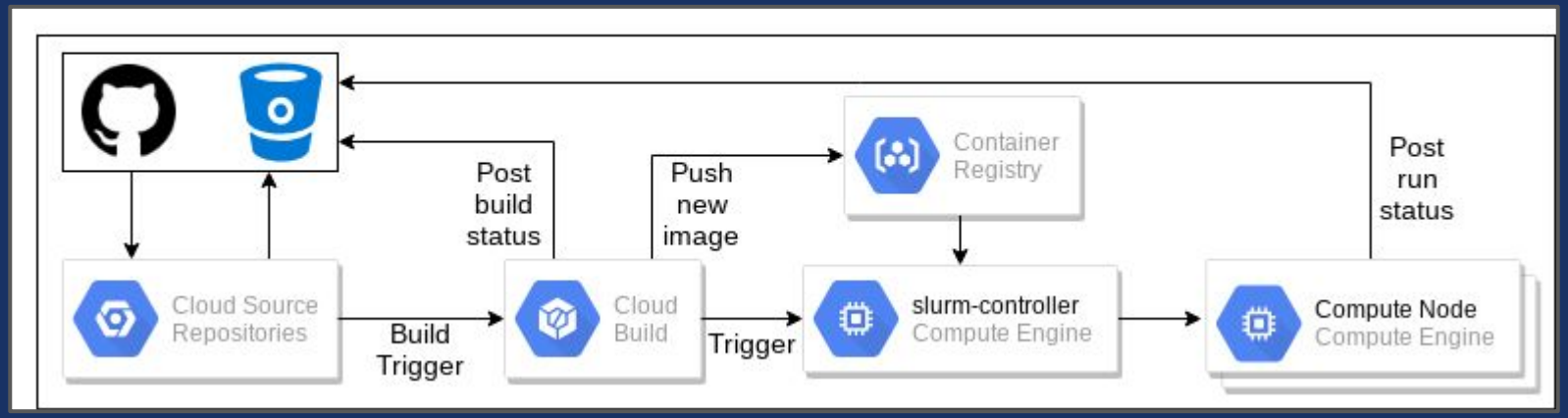
Cloud Cost Analysis



- Standard-32
- Highmem-8 + 1xP4 GPU
- Highmem-8 + 1xP100 GPU
- Highmem-8 + 1xV100 GPU
- Highmem-64 + 8xV100 GPU

FluidNumerics

HPC DevOps



FluidNumerics

A Cultural Shift

Currently, organizations maintain one cluster that supports many teams

Teams of scientists/developers are separated from the infrastructure and admin teams

Developer teams or divisions can experiment and develop on their own cluster.

This would require infrastructure and admin individuals to interact more closely with scientists and developers

FluidNumerics

Impromptu Tutorial this afternoon

At 3:45 , we'll do a tutorial where you can
spin up your own elastic slurm cluster on GCP.

https://codelabs.developers.google.com/codelabs/hpc-slurm-on-gcp/

FluidNumerics

Further Questions

Contact : joe@fluidnumerics.com

https://fluidnumerics.com

FluidNumerics