Leveraging local resources via Openstack*

Andrew Washbrook
University of Edinburgh

ScotGrid Face to Face Meeting

10th February 2017

*This is fancy title given to me by Gareth (but better than "How does this Openstack thing work again?")

Edinburgh Research Cloud

- The **Edinburgh Research Cloud*** is an *Infrastructure as a Service* Cloud for carrying out computational and digital research.
- Research Services will provide free and paid access
- **Free tier -** All academic staff and research students have access to a restricted amount of cloud resource at no cost

- Early adoption phase in late 2016
  - We were early adopters for GridPP and LSST
- General availability service announcement last week

** We'd like to invite suggestions for an appropriate name for the new cloud service. Suggestions involving the University's history or mission are particularly welcome. Please email suggestions by 9am on Monday 13th February to name_your_cloud@mlist.is.ed.ac.uk.*

Service Details

- Self-service infrastructure management via Openstack Horizon interface
- Backing Storage provided by Object storage (500TB+)
- **Sahara** available for *Big Data* cluster provisioning
- Image flavours pre-defined by RS

Cost of running on the cloud will not be more expensive than running the equivalent workload on the main Eddie 3 cluster

Flavor	vCPUs	RAM	Hard Disk	Cost per month
t1.tiny	1	512 MB	10 GB	Free
t1.small	2	1 GB	20 GB	Free

Flavor	vCPUs	RAM	Hard Disk	Cost per month
m1.small	1	2 GB	20 GB	£7.14
m1.medium	2	4 GB	40 GB	£14.29
m1.large	4	8 GB	80 GB	£28.57
m1.xlarge	8	16 GB	160 GB	£57.14

Flavor	vCPUs	RAM	Hard Disk	Cost per month
l1.small	1	6 GB	40 GB	£14.37
l1.medium	2	12 GB	80 GB	£28.73
l1.large	4	24 GB	160 GB	£57.46
l1.xlarge	8	48 GB	320 GB	£114.93

Costs for each supplied image (as of Feb 2017)

Why are we interested?

- Our original motivation arose from the provision of SL6 resources to satisfy ATLAS analysis requirements
- However this solves a growing number of limitations in our existing shared facility model

- Limited access to worker nodes for troubleshooting and package and security updates

Why are we interested?

- Our original motivation arose from the provision of SL6 resources to satisfy ATLAS analysis requirements
- However this solves a growing number of limitations in our existing shared facility model

- Limited access to worker nodes for troubleshooting and package and security updates
- Full admin access to all deployed instances
- Free to define our own set of images and to update them at our own pace

Why are we interested?

- Our original motivation arose from the provision of SL6 resources to satisfy ATLAS analysis requirements
- However this solves a growing number of limitations in our existing shared facility model

- Limited access to worker nodes for troubleshooting and package and security updates
- Full admin access to all deployed instances
- Free to define our own set of images and to update them at our own pace
- Grid specific configuration (functional accounts, CVMFS, NFS, VO-specific dependencies)

Why are we interested?

- Our original motivation arose from the provision of SL6 resources to satisfy ATLAS analysis requirements
- However this solves a growing number of limitations in our existing shared facility model

- Limited access to worker nodes for troubleshooting and package and security updates
- Full admin access to all deployed instances
- Free to define our own set of images and to update them at our own pace
- Grid specific configuration (functional accounts, CVMFS, NFS, VO-specific dependencies)
- Include configuration as part of our image build process
- Use snapshots to overlay variations on standard OS builds to suit any VO requirements
- Get standard images/containers from upstream

Why are we interested?

- Our original motivation arose from the provision of SL6 resources to satisfy ATLAS analysis requirements
- However this solves a growing number of limitations in our existing shared facility model

- Limited access to worker nodes for troubleshooting and package and security updates
- Full admin access to all deployed instances
- Free to define our own set of images and to update them at our own pace
- Grid specific configuration (functional accounts, CVMFS, NFS, VO-specific dependencies)
- Include configuration as part of our image build process
- Use snapshots to overlay variations on standard OS builds to suit any VO requirements
- Get standard images/containers from upstream
- Grid workload prioritisation has to be included into main cluster scheduler
  - Fairshare compared to other cluster users, per-VO shares

Why are we interested?

- Our original motivation arose from the provision of SL6 resources to satisfy ATLAS analysis requirements
- However this solves a growing number of limitations in our existing shared facility model

- Limited access to worker nodes for troubleshooting and package and security updates
- Full admin access to all deployed instances
- Free to define our own set of images and to update them at our own pace
- Grid specific configuration (functional accounts, CVMFS, NFS, VO-specific dependencies)
- Include configuration as part of our image build process
- Use snapshots to overlay variations on standard OS builds to suit any VO requirements
- Get standard images/containers from upstream
- Grid workload prioritisation has to be included into main cluster scheduler
  - Fairshare compared to other cluster users, per-VO shares
- Can now provision our own scheduler to fit the needs of Grid computing
- Maintain our own scheduler with our chosen VO-based ruleset

Why are we interested?

- Our original motivation arose from the provision of SL6 resources to satisfy ATLAS analysis requirements
- However this solves a growing number of limitations in our existing shared facility model

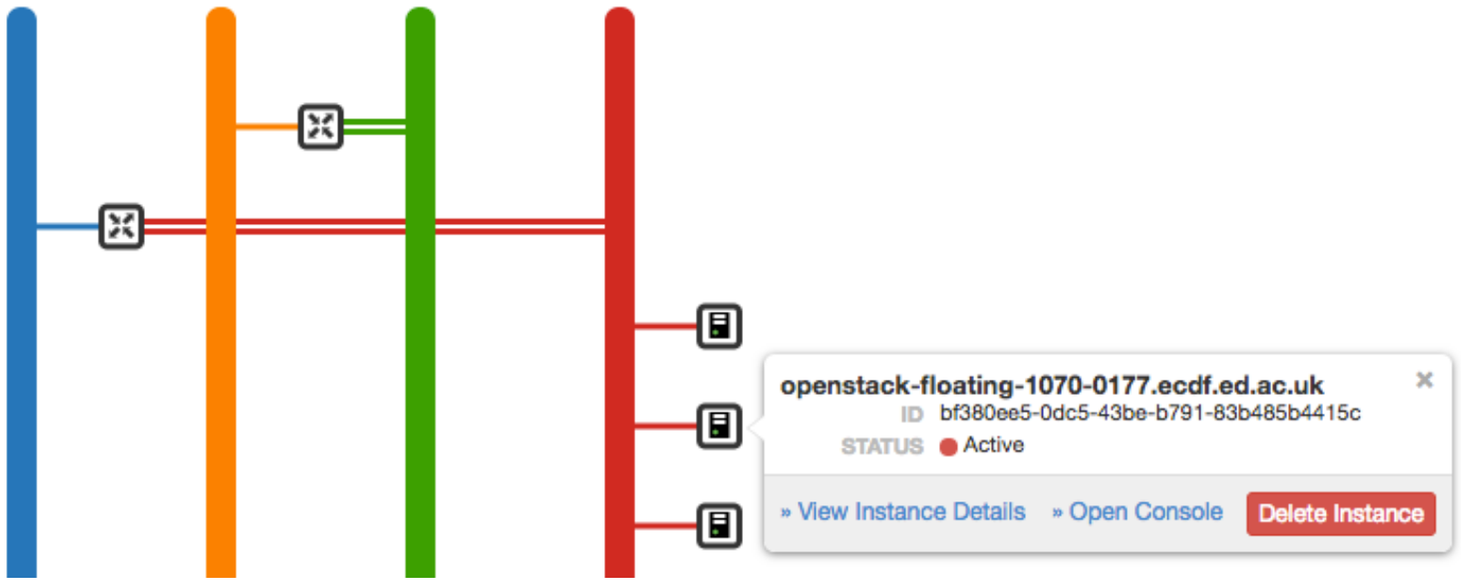
- Limited access to worker nodes for troubleshooting and package and security updates
- Full admin access to all deployed instances
- Free to define our own set of images and to update them at our own pace
- Grid specific configuration (functional accounts, CVMFS, NFS, VO-specific dependencies)
- Include configuration as part of our image build process
- Use snapshots to overlay variations on standard OS builds to suit any VO requirements
- Get standard images/containers from upstream
- Grid workload prioritisation has to be included into main cluster scheduler
  - Fairshare compared to other cluster users, per-VO shares
- Can now provision our own scheduler to fit the needs of Grid computing
- Maintain our own scheduler with our chosen VO-based ruleset
- Meter running on our ringfenced resources when sat idle

Why are we interested?

- Our original motivation arose from the provision of SL6 resources to satisfy ATLAS analysis requirements
- However this solves a growing number of limitations in our existing shared facility model

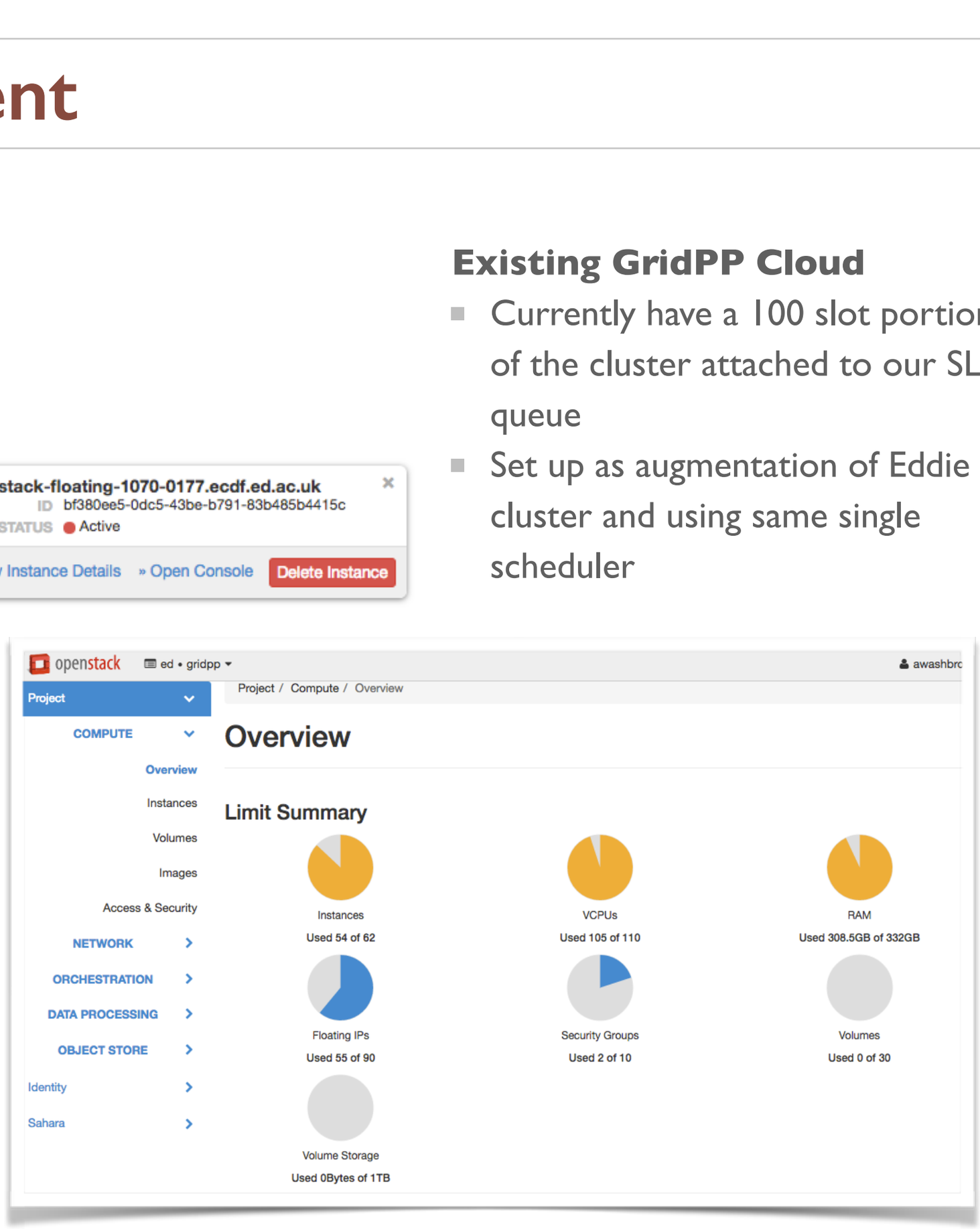
- Limited access to worker nodes for troubleshooting and package and security updates
- Full admin access to all deployed instances
- Free to define our own set of images and to update them at our own pace
- Grid specific configuration (functional accounts, CVMFS, NFS, VO-specific dependencies)
- Include configuration as part of our image build process
- Use snapshots to overlay variations on standard OS builds to suit any VO requirements
- Get standard images/containers from upstream
- Grid workload prioritisation has to be included into main cluster scheduler
  - Fairshare compared to other cluster users, per-VO shares
- Can now provision our own scheduler to fit the needs of Grid computing
- Maintain our own scheduler with our chosen VO-based ruleset
- Meter running on our ringfenced resources when sat idle
- Instances can be spun up and down depending on workload demand
- Pay as you go: if we don't use the resources we don't pay

Cloud Management



Existing GridPP Cloud

- Currently have a 100 slot portion of the cluster attached to our SL6 queue
- Set up as augmentation of Eddie cluster and using same single scheduler



Proposed Model

- For now we are being pragmatic in our approach. We want:

- To deploy a cluster we have full control of (that just happens to be hosted in a IaaS cloud)
- To run jobs from any VO without any dependency on external cloud-specific configuration or provisioning
- The ability to rapidly scale up and down cluster size in order to be more cost effective
- To not diverge from **classic** Eddie 3 processing so we can retain our opportunistic use

- Still walking through the finer details of each component
  - Appropriate choice of scheduler
  - OS API / scheduler interaction
  - Work activity heuristics
  - Set of images and configuration options
- Investigating orchestration technologies (Openstack Heat)

Other Considerations

Additional Computing Resources

- Openstack cluster only provides our paid-for resources
- Still big motivation to use opportunistic resources
- Submit jobs to Eddie 3 central scheduler as before
- Can also expand to other sources: Tier 3, HPC

- Request **KSM** to be enabled to "overclock" memory and to increase instance density
  - RS not keen if pollutes "space" of other applications
- **cgroups**-based resource management
- **Benchmarking** will vary depending on image flavour
  - This is likely to have to applied per-job to make any sense
- Information publishing will be invariably broken
  - But maybe none more than usual

Opportunistic Workload

- Job needs to have flexible lifetime, checkpointed and ideally pre-emptable
- Approaches
  - Soak up any empty resources
  - Take slots free during backfilling
  - Self-terminating jobs for higher priority workload
- ATLAS Event Service / Harvester model seems promising
  - Should not be limited to ATLAS

Summary

- Aiming to provision a GridPP cluster in the Edinburgh Research Cloud without loss of functionality to our current shared facility model

- Working within the provided service rather than creating a new product
- Not expecting this to be a model to follow elsewhere
- Many benefits including easier resource management and cost effectiveness
- New approach needs to be fully tested - requires some R&D
- Effort is fully supported and encouraged by Edinburgh Research Services
- They will use our work to feed in to solutions for other research groups

Deployment Schedule

- Aiming for an Q217 switchover
- Ambitious! At the very least run concurrently with ringfenced nodes
- If it proves unworkable we can fall back to either ringfenced (or PAYG) model on Eddie 3 cluster