EMBL European Molecular Biology Laboratory

EMBL-EBI Hybrid Cloud Pilot - Phase 1

Editors: Steven Newhouse & Conor McMenamin
May 2017

Context

Over the last 12 months, the Technical Services Cluster (TSC) at EMBL-EBI (www.ebi.ac.uk) has been exploring the use of external cloud providers to complement the infrastructure that EMBL-EBI has on site - this is termed a hybrid-cloud model. The work started within the Technology Science Integration (TSI) team[1], but following the cloud tender engaged teams across the whole of TSC, has been undertaken in four distinct phases:

- February 2016 - July 2016: A sampling of diverse cloud resource that allows the TSI team to establish sufficient information to develop a cloud benchmarking suite and to define a technical specification for an EMBL tender.
- July 2016 - September 2016: Definition, publishing and evaluation of the EMBL tender for three types of cloud services that were awarded to UKCloud (OpenStack and VMware), Microsoft Azure (MSA) and Google Cloud Platform (GCP).
- September 2016 - February 2017: Post contract negotiations which were quickly completed with UKCloud but took significantly longer with Google and Microsoft to resolve legal issues due to EMBL-EBI's intergovernmental status.
- February 2017 - May 2017: Technical exploitation of the tendered cloud resources by the Hybrid Cloud Working Group encompassing teams from across TSC which are detailed below.

Scope

A number of technical use cases within a hybrid-cloud scenario have been evaluated. These are detailed in the report but can be summarised by:

- Evaluation and deployment into production of Global Server Load Balancing (GSLB) for some EMBL-EBI web space using Amazon Web Services.
- Establishment of TSC's compute cluster environment in OpenStack and GCP using a defined and repeatable deployment recipe. MSA is in progress.
- Deployment of a local Marine Metagenomics (MMG) pipeline onto OpenStack and GCP. MSA is in progress.
- Deployment of a representative web service workload (web service application and database instance hosted on Delphix) taken from our internal VMware environment into an off-site VMware environment.

[1] https://www.ebi.ac.uk/about/people/steven-newhouse

Lessons Learnt

These technical use case of a hybrid cloud environment have led to the following lessons being learnt:

- The environments needed to support the selected internal workloads have been successfully replicated on the selected external cloud resources, but not yet on all cloud resources.
- The effectiveness of some workloads (e.g. MMG pipeline) on the cloud resources with large numbers of cores and relatively (to the number of cores) small data sets did not scale well. This is an issue with the scientific applications, not the cloud infrastructures.
- User workloads, such as the MMG pipeline, need to be adapted to make more targeted use of cloud resources which are provisioned when they are needed. This could considerably reduce the costs.
- Deployment of our internal web service hosting environment supported by VMware (web server VM, database VM, filestore) onto the VMware environment from UKCloud was achieved and the endpoint was configured to successfully receive requests as part of a web load balancer service pool. This could allow EMBL-EBI to serve web services from either inside our own data centre or from a cloud provider transparently to the end-user.
- The adoption of GSLB across the whole of EMBL-EBI's web space would require service endpoints to be referenced by a unique hostname rather than a unique URL and is now being considered on a case by case basis.
- The GCP was felt to have the most mature environment for the tracking and allocation of costs to internal subprojects. MSA was felt to have the steepest learning curve around its adoption due to the current (May 2017) incomplete support in the DevOps ecosystem around the new Azure interface.
- Documentation has been developed to provide inform team leaders at EMBL-EBI's information on how to select cloud resources, the technical issues in making applications 'cloud ready' and the procedures needed to manage cloud usage.
- The EMBL tender process was used to procure three lots of cloud computing from three different suppliers. Post-tender contract negotiation was complicated due to EMBL-EBI's privileges and immunities, and concerns around data protection and privacy which were mitigated in this phase by focusing on just public data.
- The costs between different cloud providers have not provided a significant differentiating factor compared to ease of use and ease of administration. The cost between the external cloud providers and EMBL-EBI resources currently appear to be larger, however the accuracy of the EMBL-EBI costs continues to be improved and made consistent across the internal teams. For instance, many of these costings assume full utilisation of the resource, but if that is not being achieved then the consumed units costs will be greater. In a hybrid-cloud model cheaper external costs are not necessarily a requirement, instead the ability to provision extra capacity on demand for a short period at higher cost may be more effective than having cheaper capacity purchased and in-house but idle.

Future Work

The next steps in the hybrid cloud pilot are to:

- Work with application teams to make their workloads (e.g. MMG) more performant and cost-effective when running on cloud resources. This includes requesting the right resource (e.g. virtual machine size) for the right application in the pipeline and to provision that machine for just when it is needed. This would provide sufficient machines of the right size to meet the demand and then remove these machines when they are no longer being used to manage costs. Further adaption could includes adjusting the pipelines to deal with preemptible resources (e.g. spot market) that can be reclaimed on short notice by the cloud provider but are much cheaper to use. Individual jobs in this scenario might have to be re-run if the machine was taken back by the cloud provider.
- Link the on-site LSF clusters to those in the cloud providers to explore the free movement of jobs between clusters based upon policy (e.g. data set dependency and availability in the cloud provider). Once this linkage is established production issues such as reading data over NFS from EMBL-EBI and writing to cloud based storage from the cloud based clusters will be explored. A self contained workload (applications and data relating to Uniprot's weekly release pipeline) will be run on the cloud clusters to establish a baseline performance. Work within the HelixNebula Science Cloud (http://www.helix-nebula.eu/) project is identifying some technologies that might help with this use case.
- Consider how the ability to run a web hosting environment on an external cloud provider should be embedded into TSC's operational policies so that any future deployment for planned or emergency data centre outages can be easily repeated. The adoption of a container model and a more consistent 'infrastructure as software' which would allow application requirements to be tuned and deployment to be automated has already started as part of a continuous integration and continuous deployment project.
- Continue the roll out of GSLB (provided by AWS's Route 53) across EMBL-EBI to improve service fail-over and simplify configuration and operation of the existing traffic management appliances in each data centre.
- Consider exporting costs from an external cloud provider programmatically into the Resource Usage and Accounting Portal being developed within TSC so that the external cloud provider expenditure can be shown alongside other TSC resource costs.
- Continue the work to federate EMBL-EBI's OpenStack based Embassy cloud with an external provider once technical compatibility and staff resources are aligned.
- Review and document the current bottom-up team based costings so that they are consistent with each other (e.g. in terms of staff costs, data centre costs, networking etc) and are aligned with internal cost centres so that these costs can be regularly and consistently updated over time.

The collaboration across the Technical Services Cluster within this Hybrid Cloud project has been extremely productive and has spawned a number of a new service development areas:

containers, continuous integration/delivery, and identified how our workloads could be made more portable and resource constrained. The collaboration outside of TSC with internal users has also been incredibly useful in driving through this work with real internal workloads.

Acknowledgements

Many thanks to the contributions of the following EMBL-EBI staff which were active in this project from various teams - Web Production, Systems Applications, Systems Infrastructure, Technology Science Integration and the Strategic Project Management Office - are detailed below:

- Delisa Simonovic
- Luis Gracia
- Gianni Della Torre
- Dario Vianello
- Jonathan Barker
- Russell Vincent
- Andrea Cristofori
- Salvatore Dinardo
- Alessio Checcucci
- Sunny Nanuwa
- Tomasz Nowak
- Michal Wieczorek
- Andrew Bone
- Charles Short
- Manuela Menchi
- Luis Figueria

Thanks also to the service teams that have been cooperating in this work:

- Rob Finn and the Marine Metagenomics team
- Maria Martin and the Uniprot team

Our thanks to BBSRC for supporting this work through the Large Facilities Capital Fund grant that supports EMBL-EBI's data centre environment and the equipment operated within it.

Further Information

For more information on these activities please contact steven.newhouse@ebi.ac.uk.