

ABSTRACT FORM (for new projects only)

Project Title: 901503 Feasibility Study for Applying Cloud Infrastructure to Data Analysis Problems

PI	Project Manager	Lead Scientist	Lead Engineer
Brian Schlining			

(Indicate individuals appropriate for your project. Not all projects need all four positions filled in.)

Project Description:

MBARI is continually collecting and archiving large amounts of disparate data. Some of these data sets are becoming large enough that they cannot be effectively analyzed on a single computer. Examples, of these data sets include analysis of VARS annotations, genetic sequence analysis and AVED image segmentation and classification. Cloud services and infrastructure may provide a cost-effective means to use large on-demand clusters of compute nodes to overcome analysis bottlenecks. In addition, access to on-demand computing may allow researchers to analyze data in ways that would not have been possible just a few years ago.

”Cloud services” refers to networks of computers hosted by an external company, such as Google, Amazon, Microsoft, etc. In the past year, several developments have occurred making cloud services especially attractive. These include:

1. A significant reduction in costs. For example, running twenty 8-core compute engines on Google’s Cloud for 86 hours costs about \$30 (as of June 2014).
2. There are now a variety of robust, distributed frameworks for data analysis, such as Akka (<http://akka.io/>) and Spark (<https://spark.apache.org/>). These frameworks greatly reduce the engineering effort required to develop distributed applications.
3. Development of Docker (<http://www.docker.com>), an open platform for distributed applications. This allows ‘dockerized apps’ to be assembled as part of a build process. These dockerized apps can in turn be deployed on a variety of different cloud providers so we can avoid vendor lock-in.

We are proposing a feasibility study to investigate cloud services and associated technologies as they can be applied to MBARI’s specific data analysis needs. The goals of this study are to investigate the techniques and workflows required for deploying cloud-based applications as well as to develop in-house knowledge and expertise of cloud services. As part of the feasibility study, we intend to develop a prototype application using industry best-practices that addresses a specific MBARI problem. This application will in-turn be deployed and run on a cloud service provider.

2015 Abstract Form – Internal-Only Projects (no external funding).

Deadline for abstract submission is 8:00 a.m. on June 20th.

This effort aligns with the MBARI strategic as it fosters innovation and engages in cutting-edge technological developments to spur and refresh ocean science research. It also supports MBARI's technical roadmap: Problem Statement #7: "Devise informatics tools for understanding the ocean and its inhabitants".