# Predictive Auto-scaling in the Kubernetes Cluster Manager

F. Matt McNaughton[1], S. Jeannie Albrecht[1], T. Brendan Burns[2]

[1]Department of Computer Science
Williams College

[2]Lead Engineer for Kubernetes
Google

Department Proposal Talk, 2016

# Outline I

# Outline II

- Current State
- Future

# Outline

# General

Contribute to distributed system's ability to reliably and resourcefully perform large, varying amounts of computational work.

# Outline

# Specific

We seek to maximize the sum of two metrics: Efficient Resource Utilization and Quality of Service.

# Efficient Resource Utilization (ERU)

A measure of whether an application is efficiently using the resources it is given.

# Quality of Service (QOS)

A measure of whether the application is accomplishing its stated purpose.

# Balancing ERU and QOS

Our goal is to maximize the summation of ERU and QOS. We want one of the following:

- ERU to increase and QOS to stay constant.
- ERU to stay constant and QOS to increase.
- Both!

Accomplishing these goals can have substantial real world impacts.

# Outline

# Outline

# Outline

# Outline

# Outline

# Outline

# Outline

# Outline

# Outline

# Outline

# Citations

Check out the k8s website.[1].

# Citations I

 *Kubernetes Website*. http://kubernetes.io.