Distributed and cloud computing services are increasingly built atop a preexisting infrastructure of shared services. These services have separate performance characteristics and require enough resources to support each application's service level objectives (SLOs), while preferably not wasting too many resources from overprovisioning. Changes in a service's performance are common (e.g., multiple times per day) for any number of reasons, such as from modified system configurations, hardware failures, or increased loads. Even worse, a problem in any one service can cause cascading delays across a complex web of interdependent services.

In this proposal, we describe an automated approach to mitigating such performance problems through reactive resource provisioning. When a problem occurs, we attempt to mitigate the problem in the short term by automatically assigning the right types and quantities of resources across services that can usefully apply them. Our proposed approach makes use of end-to-end request traces to determine the actual service flow and synchronicity requirements, combined with resource usage statistics to determine specific demands. This general monitoring framework is also used to discover each service's elastic scaling properties and to provide online feedback to better evaluate resource assignments.