Autoescalado basado en Aprendizaje Profundo por Refuerzo de Workflows Científicos en la Nube

Elina Pacini[1,2], Carlos Catania[1], Yisel Garí[1] and Luciano Robino[1]
[1] Facultad de Ingeniería, UNCuyo
[2] CONICET
elina.pacini@ingenieria.uncu.edu.ar, harpo@ingenieria.uncu.edu.ar, ygari@uncuyo.edu.ar, luciano.ivan.robino@gmail.com

CONTEXTO

El presente proyecto se desarrolla en el marco de la Facultad de Ingeniería dentro Laboratorio de sistemas inteligentes (LABSIN). El presente trabajo forma parte del proyecto de investigación B038-T1 que dio inicio en el mes de mayo de 2022 en el marco de los proyectos bienales de secretaria de Investigación, Internacionales y Posgrados (SIIP) de la Universidad Nacional de Cuyo.

FORMACIÓN DE RRHH

El proyecto ha permitido la formación de un estudiante de doctorado y la continuidad de las investigaciones de una becaria posdoctoral.

WORKFLOWS CIENTÍFICOS

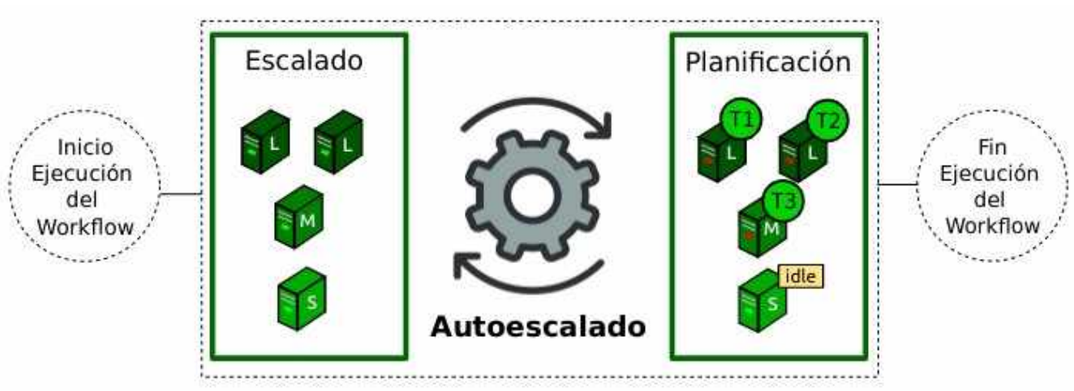
Las tecnologías workflow juegan un papel fundamental en el modelado de experimentos complejos en diferentes disciplinas científicas, facilitando la división de grandes procesos en un conjunto de componentes individualmente reutilizables y sus dependencias.
Este tipo de aplicaciones para ser ejecutadas de manera eficiente requieren de una gran capacidad de procesamiento. Además, las estructuras que describen las dependencias del workflow impactan directamente en la variabilidad de la carga de trabajo durante la ejecución.

MODELO DE COMPUTACIÓN CLOUD

Cloud facilita el acceso a los recursos computacionales necesarios para la ejecución de los workflow científicos. ofrecen recursos elásticos casi ilimitados a sus usuarios, bajo un esquema de pago por uso. Dicha eslasticidad, que posibilita el ajuste dinámico de la infraestructura, se sustenta en tecnologías de virtualización.
De esta forma, los usuarios pueden acceder a un amplio espectro de configuraciones de hardware y software, donde cada tipo de Máquina Virtual (MV) tiene un costo asociado dependiendo de sus prestaciones.

AUTOESCALADO DE WORKFLOWS EN CLOUD

Las estrategias de autoescalado explotan la elasticidad de Cloud para hacer frente a las demandas computacionales dinámicas de los workflows. Periódicamente dichas estrategias escalan la infraestructura, adquiriendo y/o terminando instancias de MVs con diferentes prestaciones. También de forma periódica, las estrategias planifican eficientemente las tareas en las instancias disponibles.
En general, las estrategias de autoescalado toman decisiones de escalado y planificación para optimizar determinados objetivos como pueden ser el tiempo de ejecución y el costo económico.



Proceso Cíclico de Autoescalado que incluye los subprocesos de escalado y planificación. L, M y S representan Mvs con diferentes capacidades computacionales. T1, T2 y T3 son tareas a ejecutar.

APRENDIZAJE PROFUNDO POR REFUERZO PARA AUTOESCALADO

El Aprendizaje por Refuerzo (AR) constituye uno de los paradigmas del aprendizaje automático y se trata de un enfoque computacional que permite a un agente aprender, mediante la interacción con el entorno, un comportamiento adecuado para el logro de un determinado objetivo. Luego, el Aprendizaje Profundo por Refuerzo (APR) combina los conceptos y algoritmos representativos del AR con las potencialidades de las RNP permitiendo manejar eficientemente espacios de estados y acciones de grandes dimensiones. De esta manera, ante entornos complejos, el APR ofrece un marco de trabajo más potente que el ofrecido por las estrategias clásicas del AR, ya que posibilita un trabajo eficiente con caracterizaciones más detalladas de dichos entornos.
El autoescalado de workflows en Cloud es un problema de toma de decisiones en un entorno estocástico complejo donde elementos como la variabilidad en el performace de la infraestructura Cloud y los patrones de carga de trabajo variables durante la ejecución de las aplicaciones son factores de incertidumbre importantes al gestionar correctamente los recursos de un entorno de computación Cloud.
Además, la heterogeneidad que caracteriza a Cloud, con una amplia gama de opciones en cuanto a tipos de MVs y modelos de precios, remarca la importancia de enfoques como el APR donde es factible el trabajo con espacios de estados y de acciones de grandes dimensiones, permitiendo trabajar con una mejor caracterización del entorno. En este sentido, la línea de investigación que se propone en el presente proyecto es novedosa y tiene mucho potencial para el desarrollo de nuevos aportes.

RESULTADOS OBTENIDOS

Durante esta primera etapa del proyecto se completaron las siguientes actividades:

a) Actualización del estado del arte sobre estrategias basadas en APR aplicadas al problema de autoescalado en Cloud. Dicho relevamiento complementa el trabajo previo realizado en [1].
b) Diseño y desarrollo de un autoescalador basado en AR. Algunos resultados preliminares fueron publicados en [2].
c) Actualmente se está trabajando en una extensión del trabajo publicado en [2], incorporando nuevos algoritmos tanto basados en AR como APR. Se espera publicar los resultados durante el transcurso del presente año.

REFERENCIAS

[1] Yisel Garí, David A. Monge, Elina Pacini, Cristian Mateos, and Carlos García Garino. Reinforcement learning-based application autoscaling in the cloud: A survey. Engineering Applications of Artificial Intelligence, 102:104288, 2021

[2] Yisel Garí, David A. Monge, and Cristian Mateos. A q-learning approach for the autoscaling of scientific workflows in the cloud. Future Generation Computer Systems, 127:168–180, 2022