

# Resumen

Los flujos de trabajo científicos son comúnmente usados para modelar aplicaciones en la e-Ciencia. En este modelo de programación, las aplicaciones científicas son descritas como un conjunto de tareas que tienen dependencias entre ellas. En las últimas décadas, la ejecución de flujos de trabajo científicos se ha llevado a cabo con éxito en las infraestructuras de computación disponibles (supercomputadores, clústeres y grids) utilizando programas conocidos como Sistemas de Gestión de Flujos de Trabajo (SGFT), los cuales distribuyen la carga de trabajo en dichas infraestructuras.

No obstante, dado que cada infraestructura de computación posee su propia arquitectura y que, además, cada aplicación científica explota eficientemente una de las infraestructuras, se hace necesario organizar, de alguna manera, cómo estas aplicaciones son ejecutadas. Con el fin de abordar el procesamiento de grandes cantidades de datos, los SGFT necesitan aprovechar al máximo todos los recursos de computación y almacenamiento disponibles.

Originalmente, las aplicaciones científicas se ejecutaban masivamente en infraestructuras de computación de altas prestaciones (tales como supercomputadores y clústeres) y los grids. Sin embargo, en la última década, ha aparecido un nuevo paradigma conocido como Computación en la Nube, el cual ha permitido el uso de infraestructuras bajo demanda para complementar o incluso sustituir infraestructuras locales. No obstante, este hecho plantea nuevas cuestiones, tales como la integración de recursos híbridos o el compromiso entre la reutilización de infraestructuras y la elasticidad, todo ello teniendo en cuenta la eficiencia en coste.

La principal contribución de esta tesis es una solución orientada a la gestión de flujos de trabajos, explotando las capacidades de los orquestadores de Computación en la Nube para desplegar recursos bajo demanda, en función de la carga de trabajo, y para combinar proveedores de Computación en la Nube heterogéneos (públicos y privados) e infraestructuras tradicionales (supercomputadores y clústeres) que minimicen el tiempo de respuesta. El objetivo de la tesis no es aportar otro SGFT más a la larga lista que puede hallarse en el estado del arte, sino demostrar los beneficios que se obtienen al integrar los orquestadores de Computación en la Nube cuando se pretende ejecutar flujos de trabajo complejos. La tesis muestra experimentos con diferentes configuraciones y múltiples recursos heterogéneos, haciendo uso de un flujo de trabajos de genómica comparativa, llamado Orthosearch, con el fin de migrar carga de trabajo que hace uso intensivo de memoria a infraestructuras públicas mientras se ejecuta localmente el resto de bloques del experimento. El tiempo de ejecución y el coste de los experimentos son detallados, junto con buenas prácticas que es aconsejable seguir a la hora de abordar escenarios similares.