

# Calendarización de flujos de trabajo en cómputo en la nube

Fernando Aguilar Reyes

10 de noviembre de 2013

## 1. Análisis de antecedentes: Revisión inicial de literatura

Entendemos que un flujo de trabajo, o *workflow* en inglés, es un conjunto de pasos que modelan la ejecución de un proceso [4]. Un ejemplo de un flujo de trabajo sería la anotación de proteínas, el cual consiste en identificar ciertas partes de la proteína y poner una etiqueta que describa la función de dicho componente [5].

Por lo general, los flujos de trabajo tienen una complejidad computacional que hace prohibitivos ejecutarlos en una sola computadora. De acuerdo a Buyya et al., diversos enfoques se han aplicado para distribuir la ejecución de un flujo de trabajo entre varias computadoras. Los enfoques de cómputo más importantes son los clusters, los grids y las nubes [3]. A continuación, explicaremos cada uno de los enfoques.

- Los *clusters* son sistemas distribuidos, paralelos, compuestos de varias computadoras que son vistas como un único recurso de cómputo [3]. Un ejemplo de un cluster es la instalación de la Universidad Autónoma Metropolitana, campus Iztapalapa, llamado *Atizaloo*, compuesto por 270 nodos de cómputo, equipados cada uno con 2 procesadores Intel Xeon Quad-Core y 16GB en RAM; los nodos están conectados entre sí por medio de switches Ethernet e Infiniband. El cluster también cuenta con un sistema de almacenamiento distribuido basado en Lustre. La conexión con Internet se administra con el nodo maestro llamado *Atizaloo*. La capacidad de cómputo del cluster *Atizaloo* es de 18.4 teraFLOPS [2].
- Los *grids* son sistemas distribuidos, paralelos, compuestos de computadoras autónomas y geográficamente distribuidas que pueden trabajar en conjunto o de manera independiente de acuerdo a los objetivos, políticas y mecanismos de uno o varios administradores del sistema, es decir, un grid puede ser compartido entre varias instituciones [3]. El cluster *LANCAD* es un buen ejemplo, pues une el cluster *Atizaloo* de la UAM, el cluster

de la UNAM *KamBalam*, y el cluster *Xiuhcoatl* del CINVESTAV por medio de una red de fibra óptica instalada en las estaciones del Sistema de Transporte Colectivo Metro. La suma de la potencias reales de cada cluster robusto de cálculo del grid es de 48.55 teraFLOPS [1].

- Las *nubes* (clouds) son sistemas distribuidos, paralelos, compuestos de computadoras o máquinas virtuales interconectadas que son aprovisionadas para usarse como uno o varios recursos de cómputo, de acuerdo a un contrato de nivel de servicio acordado entre el proveedor de la nube y el cliente [3]. Empresas nuevas y existentes proveen servicios de cómputo en la nube, tales como GoGrid, Rackspace, Amazon, Microsoft, IBM, Oracle, entre otras. La forma en que operan es muy sencilla: se paga cierta cantidad por utilizar servicios de cómputo o almacenamiento durante determinado tiempo. Así, los clientes no tienen que invertir grandes cantidades de dinero para contar con una gran infraestructura como lo son los clusters y los grids.

El *cloud computing* o enfoque de cómputo en la nube está tomando mucho interés tanto por la industria como por la comunidad científica, porque hace accesible una gran cantidad de recursos computacionales con cantidades razonables de presupuesto.

## Referencias

- [1] Cluster híbrido de supercómputo — xiuhcoatl — cinvestav. <http://clusterhibrido.cinvestav.mx/>. Consultado el 10 de noviembre de 2013.
- [2] Laboratorio de supercómputo y visualización en paralelo. <http://supercomputo.izt.uam.mx/infraestructura/aitzaloa.php>. Consultado el 10 de noviembre de 2013.
- [3] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic. Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation computer systems*, 25(6):599–616, 2009.
- [4] J Octavio Gutierrez-Garcia and Kwang Mong Sim. Agent-based cloud workflow execution. *Integrated Computer-Aided Engineering*, 19(1):39–56, 2012.
- [5] Angela O’Brien, Steven Newhouse, and John Darlington. Mapping of scientific workflow within the e-protein project to distributed resources. In *UK e-Science All Hands Meeting*, pages 404–409, 2004.