Diseño del Sistema HPC-Linux-Clúster del IGP

Oscar Santillán, Huber Gilt, Augusto Ingunza, Kobi Mosquera e Ivonne Montes
Instituto Geofísico del Perú

En la actualidad, ampliar el conocimiento de sistemas complejos, por ejemplo, los fenómenos geofísicos naturales y/o de los causados por la intervención humana (antropogénicos), que pueden generar importantes pérdidas socioeconómicas, implica la utilización de la técnica de modelado numérico como herramienta. Esta técnica contempla el comportamiento de un sistema expresado a través de ecuaciones matemáticas que serán resueltas en diferentes escalas de espacio y tiempo. Sin embargo, dichas expresiones matemáticas difícilmente pueden ser resueltas analíticamente y requieren computadoras para obtener resultados numéricos.

A nivel mundial con el desarrollo de la computación y las tecnologías informáticas que permiten tener dispositivos de mayor rendimiento, alta velocidad y bajo costo, el modelado numérico de procesos geofísicos emplean computadoras superiores a las convencionales capaces de resolver miles de millones de cálculos por segundo y por largos periodos de tiempo, debido a los niveles de complejidad que la investigación científica ha encontrado. El Instituto Geofísico del Perú (IGP), desde 1998, ha sido pionero en el uso de esta tecnología para la simulación computacional del comportamiento de la atmósfera y el océano. Sin embargo, cada año aumentan las exigencias en cuanto a las características y complejidad de los modelos usados, lo que requiere incrementar la potencia de cómputo (Segura et al., 2014).

El IGP ha implementado recientemente en su Laboratorio de Dinámica de Fluidos Geofísicos Computacional una nueva infraestructura computacional de alto rendimiento que, desde marzo de 2016, está a disposición de la comunidad científica y académica del país (Montes et al., 2016).

En este artículo se presentan algunas características y pautas que fueron seguidas para la implementación de la infraestructura computacional de alto rendimiento que, en adelante, será denominado **HPC-Linux-Clúster.**

Arquitectura del sistema HPC-Linux-Clúster

El HPC-Linux-Clúster implementado en el IGP es de tipo clúster, que puede ser definido como un aglomerado de equipos/servidores (unidades de procesamiento) que están interconectados y configurados para trabajar de manera conjunta. Este diseño es escalable, ya que se pueden añadir más unidades de procesamiento y su interconexión permite incrementar la velocidad de procesamiento mediante la paralelización, que consiste en dividir una tarea entre varios procesadores trabajando simultáneamente.

La aglomeración de las unidades de procesamiento puede ser definido de diferentes formas. La utilizada para el HPC-Linux-Clúster (Figura 1) emplea la tecnología Beowulf (Sterling et al., 1995), cuya arquitectura considera como hardware un nodo maestro o servidor principal, 20 nodos de cómputo conectados entre sí mediante una red InfiniBand de alta velocidad (56 Gigabits por segundo), y mediante una red Ethernet para la administración operativa.

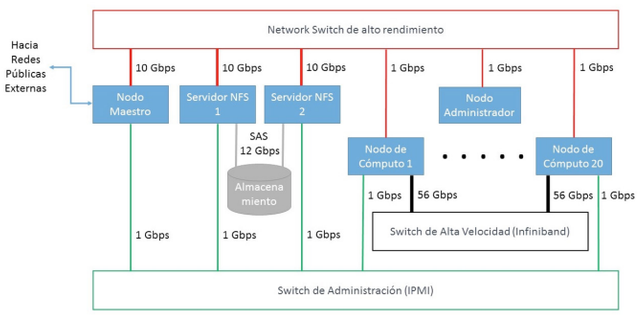

Figura 1. Configuración del Hardware del Sistema HPC-Linux-Clúster.

Estos elementos del hardware junto a un nodo administrador, ejecutan el sistema operativo libre Linux y softwares relacionados a cada nodo dependiendo de su función. Asimismo, se requiere de un sistema de clusterización para la ejecución de tareas científicas. También, existe una red interna

o privada de alto rendimiento, encargada de facilitar que todos los nodos computacionales puedan disponer del espacio físico en disco, a través del protocolo NFS (Network File System), en nuestro caso instalado en contingencia, que interactúan con un equipo de almacenamiento externo.

Con excepción del nodo maestro y el nodo administrativo, los nodos de cómputo no se encuentran equipados con un disco duro físico, ya que son encendidos a través de la red interna. Con el uso de la solución de software libre tipo DRBL (Diskless Remote Boot for Linux o inicio remoto sin disco para Linux). Para el HPC-Linux-Clúster del IGP, se usa una versión reducida del software BProc (Beowulf Distributed Process Space).

También forman parte del Sistema los diferentes equipos de comunicaciones de mediano y alto rendimiento, que permiten una adecuada administración y uso de los diferentes recursos de memoria, almacenamiento y de cómputo.

Cada nodo de cómputo y el nodo maestro están equipados con dos procesadores Intel Xeon E5-2680 de 3.5 GHZ, cada uno con 12 cores o núcleos, haciendo un total de 24 núcleos por nodo; logrando una sumatoria total de 480 núcleos (considerando los 20 nodos computacionales). Estos equipos también poseen cada uno 8 módulos de memoria tipo 2133 MT/s RDIMM de 16 Gigabytes (GB), totalizando 128 GB de memoria RAM por cada nodo.

El almacenamiento del nodo principal permite ofrecer a los usuarios del HPC-Linux-Clúster un espacio aproximado de 15 Terabytes (TB) de capacidad, sumado a un equipo de almacenamiento externo que contiene 60 discos tipo NLSAS de 6 TB configurados en un arreglo tipo RAID 6, permitiendo disponer de alrededor de 294 TB de almacenamiento externo.

Especificaciones del Software asociado al sistema HPC-Linux-Clúster

La instalación del software asociado al HPC-Linux-Clúster es un proceso de integración de varios sistemas de software, tal como se muestra en la Figura 2. Esta integración incluye un sistema operativo, programas compiladores, un sistema de red o de networking, un sistema de paralelización, un sistema de organización y planificación de trabajos, opcionalmente un sistema de monitoreo y, finalmente, un paquete de aplicaciones variado.

El sistema operativo seleccionado para el HPC-Linux-Clúster está basado en el denominado Open SUSE Linux LEAP versión 42.1, el cual se encuentra instalado tanto en el servidor principal como en el nodo administrativo.

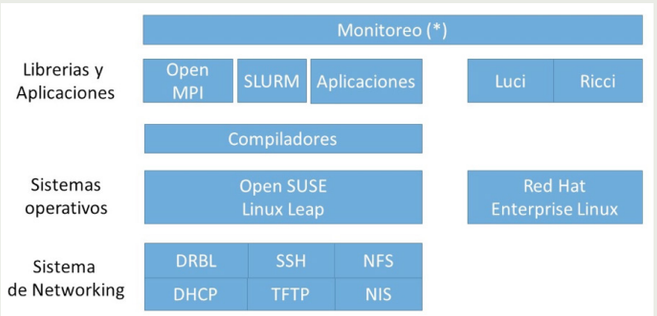

Figura 2. Estructura del Software del Sistema HPC-Linux-Clúster.

Adicionalmente, a los compiladores GNU (gcc, g++ y gfortran) instalados por defecto, se han incluido los compiladores de Intel C y C++ (icpc e icc) y Fortran Compiler 16 (ifort) que forman parte del software Parallel Studio; que instala importantes librerías para el modelamiento numérico y matemático.

El Sistema de paralelización elegido está basado en el modelo de programación denominado Open MPI (Open Source Message Passing Interface), que combina la experiencia, tecnología y recursos de toda la comunidad HPC (High Performance Computing), para obtener la mejor librería disponible para desarrolladores e investigadores científicos; la cual soporta múltiples trabajos y un procesamiento tolerante a fallos.

Para realizar el manejo y la administración de los múltiples nodos computacionales disponibles, y al mismo tiempo para manejar los distintos trabajos así como los recursos de hardware del Sistema, se hace uso del software denominado SLURM (Simple Linux Utility Resource Management); el cual organiza, planifica, contabiliza, asigna y define prioridades en la ejecución de los distintos trabajos para los diferentes usuarios, el cual permite a su vez administrar los resultados.

Para permitir que los diferentes usuarios del HPC-Linux-Clúster accedan al espacio en disco disponible en el almacenamiento externo, se ha instalado un mini clúster en los dos servidores NFS. Para ello, el sistema operativo seleccionado en dichos servidores es la versión 6.7 del denominado RHEL (Red Hat Enterprise Linux), sobre el cual se encuentra instalado y configurado los softwares llamados Luci y Ricci, que constituyen el Sistema

Oscar Santillán, Huber Gilt, Augusto Ingunza, Kobi Mosquera e Ivonne Montes
Instituto Geofísico del Perú

mini clúster antes señalado. Dicho Software permite acceder a los datos e información a través de un sistema de alta disponibilidad tolerante a fallos, ya que constantemente monitorea, controla y vigila la actividad o pasividad de ambos servidores NFS.

Completan el Sistema HPC-Linux-Clúster importantes equipos y sistemas de apoyo alterno, tales como los sistemas de manejo y administración de energía eléctrica (UPS, Generador, tableros eléctricos, sistemas de protección), los sistemas de aire acondicionado de precisión y de confort, una infraestructura física, de redes y comunicaciones, entre otros.

Todo lo antes descrito de forma resumida, constituye en gran parte lo que se conoce como el Sistema Computacional de Alto Rendimiento HPC-Linux-Clúster, que se encuentra a disposición de la comunidad científica y académica desde el año 2016. Para mayor información visite http://scah.igp.gob.pe/laboratorios/dfgc.

Referencias

Abdelbaky, M. (2012) A Framework for Enabling High-End High Performance Computing Resources as a Service. New Brunswick Rutgers, 22-28.

Chang, C.-T. Y. y. Y.-C. (2002) An Introduction to a PC Cluster with Diskless Slave Nodes. Tungai Science, 4, 25-46.

Montes, I., et al. (2016) Sistema computacional de alto rendimiento para la simulación de fluidos geofísicos HPC-Linux-Clúster.. Boletín Técnico "Generación de modelos climáticos para el pronóstico de la ocurrencia del Fenómeno El Niño", Instituto Geofísico del Perú, 9-10.

Open MPI, s.f. A High Performance Message Passing Libray., [En línea] Available at: https://www.open-mpi.org

Robert Van Engelen, L.W., et al. (1996) Automatic Code Generation for High Performance Computing in Environmental Modeling. In Proceedings of the 1996 EUROSIM Int'l Conf., June 10–12, 1996, Delft, The Netherlands, 421–428.

Segura, B., et al. (2014) Evolución del Sistema Computacional de Alto Rendimiento en el IGP para un mejor pronóstico y estudio de los fenómenos climáticos. Boletín Técnico "Generación de modelos climáticos para el pronóstico de la ocurrencia del Fenómeno El Niño", Instituto Geofísico del Perú, 1 Noviembre, 8-9.

Sterling, T. (2002) Chapter 1 - Introduction and Chapter 2 - An overview of Cluster Computing. s.l.:MIT Press, 1-29.

Sterling, T., et al. (1995) Beowulf: A Parallel Workstation For Scientific Computation, In Proceedings of the 24th International Conference on Parallel Processing.

Tsung-Lung, L. (2011) A Practical Guide to Building High-Performance Computing Clúster. s.l.:National Chia-Yi University, 3-22.

University of Maryland, s.f. Division of Information Technology, High Performance Computing. [En línea]
Disponible: http://www.glue.umd.edu/hpcc/help/slurm-vs-moab.html

Wikipedia, s.f. Network Time Protocol (NTP). [En línea]
Disponible: https://es.wikipedia.org/wiki/Network_Time_Protocol