Implementación uso y distribución de aplicaciones para física de altas energías en entornos colaborativos.

Umberto Cotti y Arnulfo Zepeda

1. Resumen

Se presentan los resultados del diseño y la construcción del “cluster” [1] para cálculo científico realizado con computadoras de escritorios y configurado exclusivamente con software libre y de código abierto, así como los primeros valores de la caracterización del desempeño del mismo.

2. Introducción y antecedentes

Para la realización de cálculos científicos en distintas áreas, se requiere de una capacidad de cómputo que por lo general una computadora personal no puede ofrecer. Con el fin de poder adquirir una capacidad de cómputo suficiente, a un costo relativamente bajo y utilizando como recursos básicos a las computadoras personales de escritorio, se ha venido implementando la arquitectura de “commodity cluster” que es un sistema de cómputo local que comprende un conjunto de computadoras independientes y una red interconectándolas.

En el caso particular de nuestro grupo de trabajo se requería de un “cluster” para el análisis de datos sobre rayos cósmicos de alta energía adquiridos en el Observatorio Pierre Auger [2] situado en Argentina de la Colaboración Internacional en la que la U.M.S.N.H. participa. Para resolver este problema se investigaron y analizaron varios esquemas posibles de configuración de hardware y software y realizaron las primeras pruebas que, con los recursos de los que se disponía en la Universidad, nos permitió llegar al diseño y a la construcción de nuestro primer “cluster” local experimental. Local en el sentido de que todos sus subsistemas componentes son supervisados dentro de un solo dominio administrativo que reside en un solo cuarto y es administrado como un solo sistema de cómputo. Los nodos que constituyen al “cluster” son computadoras manufacturadas comercialmente, capaces de operar de manera completamente independiente. El principal beneficio que el diseño de “commodity clusters” [1] ofrece con respecto los sistemas convencionales de cómputo masivamente paralelo se deben a que el número de nodos, capacidad de memoria por nodo, número de procesadores por nodo, y topología de interconexión pueden ser fácilmente modificados o aumentados como lo dicten la oportunidad o la necesidad sin pérdida de inversión previa.

3. Objetivos

Entre los objetivos en la construcción del “cluster”, se tenía el de alcanzar una capacidad de cómputo mayor a la ofrecida por una computadora de escritorio

o un conjunto de computadora no interconectadas. Una aproximación a una solución de incremento en capacidad de cómputo se vio como la adquisición de equipos tipo estaciones de trabajo, sin embargo esta solución era costosa y la capacidad de cómputo buscada no se cubría.

Otro de los objetivos fue poder construir una plataforma que permitiera el uso de herramientas de cómputo de alto rendimiento para física, modelado y animación 3-D, ya existentes en plataformas de escritorio y llevarlas al ambiente de computo distribuido/paralelo, de tal manera que el trabajo que comúnmente realiza una computadora se pudiera realizar en un arreglo de 5 equipos de cómputo (actualmente).

Por último se proyectaba poner a disposición de la comunidad científica, en especial del área de física, las herramientas computacionales y la capacidad de cómputo suficientes para realizar cálculos complejos en un solo centro de cómputo que ofreciera todas las herramientas necesarias. Además de esto, también el hacer extensivo el uso y aplicación de estas herramientas en otras áreas del conocimiento, tal y como se hace actualmente con las áreas de biofísica, física médica y química.

4. Materiales y métodos

Los materiales de los que se dispuso para el diseño y la construcción fueron exclusivamente los que la institución pudo ofrecer al momento de empezar el proyecto, se procuró no adquirir “hardware*”* y como regla solo se utilizó software libre. Los recursos humanos involucrados en el proyecto fueron un profesor y dos estudiantes todos del área de física y matemáticas. Para el diseño de la distribución del “hardware*”* este fue dictado por los recursos que ya estaban disponibles. Con relación al software se procuró escoger la configuración que permitiera mayor flexibilidad de expansión o reducción y sobre todo la compatibilidad con otros sistemas similares actualmente en función en particular el de la Facultad de Ciencias de la Universidad Autónoma de Puebla con la que se pretende en una segunda etapa configurar un “grid” que comparta los dos “clusters”.

5. Resultados obtenidos o esperados

El “cluster” construido [3] quedó conformado por 5 computadoras personales de escritorio con las siguientes características:

- Modelo: HP Compaq dc5750 Microtower
- CPU: AMD Athlon(tm) 64 X2 Dual Core Processor 5600+, 2,800MHz
- Disco: ATA, Hitachi HDS72168, 74GB
- Memoria: DIMM Synchronous 667 MHz (1.5 ns), 2GiB (2 módulos de un GiB)
- Tarjeta red: NetXtreme BCM5755 Gigabit *ethernet* PCI, Broadcom Corporation
- Sistema operativo: Debian GNU/Linux testing "Lenny" - Official

Snapshot amd64 CD Binary-1 20080314-00:31

- Software exclusivo para funcionamiento y del “cluster”:
  - PVM (3.4.5-11), Torque PBS 2.3.0, Maui 3.2.6p19, MPI versión Open (MPI 1.2.7~rc2-2), Ganglia 3.1.1 (Wien), BLAS

El “cluster” así configurado puede realizar cálculos de manera distribuida cuando se requiere repetir el mismo conjunto de operaciones una gran cantidad de veces para simular algún proceso, en este caso se utilizan el administrador de recursos Torque [4] y al planificador de tareas Maui [5]. Una computadora actúa como cabeza del “cluster” y el resto de las computadoras actúan como nodos de trabajo. También se pueden realizar cálculos de manera paralela a través del “Parallel Virtual Machine” (PVM) [6] y cuando se utiliza este esquema no se distingue entre las computadoras y las cinco realizan cálculos.

La red que permite la comunicación entre las computadoras del “cluster” es de tipo “ethernet”. Cada nodo se conecta a un solo concentrador de red de velocidad 10/100, del tipo comúnmente usado en hogares u oficinas. Este tipo de configuración permite ser expandido o reducido con mucha flexibilidad. De manera independiente del “hardware” del que se disponga es posible anexar nuevas computadoras que hayan sido configuradas con el mismo software.

En la construcción del “cluster” la parte relativa a la instalación de los sistemas operativos y a la conexión física de los equipos se realizó de manera estándar. En la configuración de la red y de los servicios se presentaron dificultades para la configuración de nombres en el archivo /etc/resolv.conf y posiblemente hubiera sido mejor usar un servidor DNS. La autenticación de SSH se realiza por medio de llaves de usuario, por lo que cada usuario debe generarlas y distribuirlas, por lo tanto usar llaves de autenticación basadas en “host” podría ser una mejor opción.

El “cluster” así construido tiene un poder de cálculo teórico que se puede caracterizar con el "pico del rendimiento teórico" (Rpeak) [7], esto es un parámetro calculable tanto para los procesadores de manera individual como en su conjunto. Está determinado mediante el conteo del número de operaciones (adiciones y multiplicaciones) en punto-flotante (a precisión máxima) que pueden ser completadas durante un ciclo del procesador. El cálculo de Rpeak de un equipo ó conjunto de equipos de cómputo, está dado por la expresión:

Rpeak = [CPUs] x [tasa de velocidad de reloj de CPU (GHz)] x [no. de operaciones de punto flotante por ciclo]

Donde CPUs es una cantidad que se utiliza en el caso de un sistema genérico de muchos procesadores y se define como el [número de procesadores] x [numero de cores por procesador].

En nuestro caso construimos un “cluster” con 5 equipos de cómputo, cada uno con 1 procesador y cada procesador con 2 Cores. Dado que el procesador

AMD Athlon(tm) 64 X2 Dual Core Processor 5600+ realiza 3 operaciones de punto flotante por ciclo, el valor de Rpeak que obtuvimos es de:

Rpeak = [10 CPUs] x [3 GHz] x [3 operaciones de punto flotante por ciclo] = 90 millones de operaciones de punto flotante por segundo = 90 Gflops = 90 Gflop/ sec

Con este límite teórico en mente y conscientes de que no existe una prueba de rendimiento universal que caracterice a un "cluster" de manera absoluta, sino que existen pruebas específicas, dependiendo del tipo de aplicaciones de interés; decidimos realizar una prueba clásica que es la que se utiliza para determinar la famosa clasificación de los 500 "clusters" [8] con mayor poder de cómputo en el mundo. Esta prueba denominada "High-Performance Linpack Benchmark for Distributed-Memory Computers" (HPL) [9] consta de un software que resuelve un sistema aleatorio denso de ecuaciones lineales en aritmética de doble precisión (en este caso ocupando 64 bits) en computadoras de memoria distribuida. El HPL provee un programa de pruebas y medición de tiempo para cuantificar la precisión de la solución obtenida, al igual que el tiempo que le tomó calcularla. El mejor desempeño que este software puede detectar en un sistema depende de una gran variedad de factores. Sin embargo, con algunas suposiciones restrictivas sobre la red de interconexión de las computadoras, el algoritmo usado por HPL y su implementación adjunta son escalables en el sentido de que su eficiencia paralela se mantiene constante con respecto al uso de memoria por procesador. HPL requiere además de la disponibilidad en el sistema de una implementación de la interfaz de transferencia de mensajes (MPI) [10], una implementación de los subprogramas de "Basic Linear Algebra Subprograms" (BLAS) [12] o de la "Vector Signal Image Processing Library" (VSIPL) [12].

HPL resuelve un sistema de ecuaciones lineales de orden N, siendo N el número de renglones y columnas de la matriz cuadrada de coeficientes generada aleatoriamente. Para obtener el máximo desempeño del sistema el valor de N debe de ser el máximo que permita ser almacenado en la memoria disponible y no debe de rebasar esta capacidad para no generar "swap".

Gracias a la topología con la que se construyó el cluster es posible utilizarlo de manera completa haciendo trabajar los 5 nodos o de manera parcial utilizando solamente unos de ellos. Realizamos 5 pruebas distinta activando 1, 2, 3, 4 y 5 computadoras y modificando el valor de N dependiendo de la memoria disponible en cada caso según la siguiente expresión:

$$N = \sqrt{\frac{M * (1024^2)}{8}}$$

Donde N es el orden de la matriz y M el total de la memoria de la que dispone el "cluster" en Megabytes. N es un valor entero adimensional, la memoria total se expresa en bytes para que al dividirla sobre 8 indique el total de números de doble precisión que caben en M. Como debe ser cuadrada, su raíz produce el

tamaño del lado. A este resultado lo ajustamos restandole aproximadamente un 20% del valor para no saturar completamente la memoria y dejar espacio para operaciones del sistema operativo.

Tabla 1. Valores de N óptimos para la medición del rendimiento del “cluster”, en función del número de computadoras que lo componen.

Nodos	Cores por nodo	Cores totales del cluster	Memoria por nodo en MB	Memoria total del cluster en MB	Valor de N	Valor de N ajustado
1	2	2	2,048	2,048	16,384	13,107
2	2	4	2,048	4,096	23,170	18,536
3	2	6	2,048	6,144	28,378	22,702
4	2	8	2,048	8,192	32,768	26,214
5	2	10	2,048	10,240	26,636	29,309

El rendimiento obtenido para el caso de los 5 nodos es del orden de los 20 Gflop/sec y es relativamente más pequeño del calculado teóricamente, esto se debe probablemente al sistema de comunicación que tenemos implementado para la red que no es de alta velocidad y es posible que esté perjudicando el desempeño. Estamos investigando como estas características impactan sobre el rendimiento.

Los resultados de las pruebas realizadas se resumen en la siguiente gráfica:

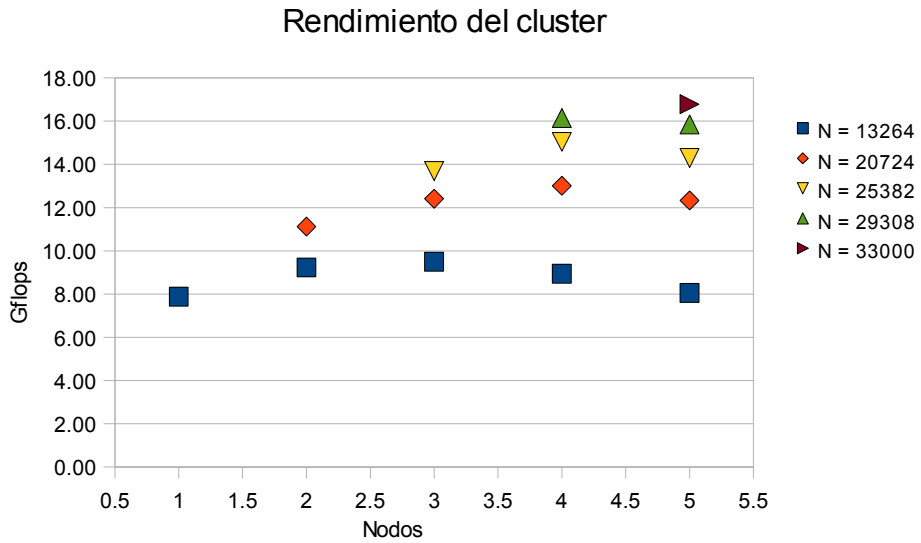

Fig. 1. Resultados de las pruebas de eficiencia para un “cluster” de 1, 2, 3, 4, 5 nodos al variar el valor de N. Se aprecia como para un valor de N dado, se alcanza un límite en la eficiencia que no cambia aunque aumente el número de nodos.

6. Conclusiones

Utilizando recursos de hardware disponible fue posible construir y caracterizar un “cluster” para cómputo científico que permitirá ahora, en la segunda etapa del proyecto, realizar los cálculos científicos requeridos para el estudio de fenómenos de física de rayos cósmicos de ultra alta energía. Al realizar la construcción del “cluster” se adquirió la experiencia necesaria para optimizar su configuración y su desempeño. Los estudiantes de licenciatura involucrados en el proyecto pudieron poner en práctica varios de los conocimientos adquiridos a lo largo de su carrera en la licenciatura en físico matemáticas. El “cluster” construido asienta las bases para el siguiente paso que será la ampliación del mismo, ahora con hardware especifico que satisfaga las necesidades detectadas.

7. Referencias bibliográficas

[1] M. Baker “Cluster Computing White Paper”, http://arxiv.org/abs/cs/0004014v2
[2] http://www.auger.org
[3] http://sirin.ifm.umich.mx/ganglia
[4] http://www.clusterresources.com/pages/products/torque-resource-manager.php
[5] http://www.clusterresources.com/pages/products/maui-cluster-scheduler.php
[6] http://www.csm.ornl.gov/pvm/pvm_home.html
[7] J. Dongarra, “Performance of Various Computers Using Standard Linear Equations Software”, Computer Science Technical Report Number CS-89. 1985
[8] http://www.top500.org/
[9] J. Dongarra, “Performance of Various Computers Using Standard Linear Equations Software, (Linpack Benchmark Report)”, Computer Science Technical Report Number CS-89-85. 2008
[10] http://www-unix.mcs.anl.gov/mpi/
[11] C. L. Lawson, R. J. Hanson, D. Kincaid y F. T. Krogh, “Basic Linear Algebra Subprograms for FORTRAN usage”, ACM Trans. Math. Soft., 5 (1979), pp. 308--323.
[12] http://www.vsipl.org