Cluster Modular Autocontenido

Gabriel E. Arellano, Leonardo J. Hoet
Departamento Ingeniería en Sistemas de Información
Facultad Regional Concepción del Uruguay
Universidad Tecnológica Nacional
[arellanog,hoetl]@frcu.utn.edu.ar

RESUMEN

Desde hace años es común en organizaciones educativas y de investigación el desarrollo de actividades que requieren grandes capacidades de cálculo. La solución más común a estas necesidades es la compra o construcción de un cluster HPC (High Performance Computing), pero construir un cluster acarrea problemas y costos inesperados problemas al momento de utilizarlo y mantenerlo en operaciones. Nuestro proyecto busca desarrollar y documentar técnicas para construir un cluster HPC que sea fácilmente ampliable y que minimice los problemas que normalmente acarrea la operación del mismo. Los resultados preliminares son alentadores, en este punto podemos asegurar un ahorro de energía de hasta un 27%, un ahorro de espacio de hasta el 30%, y un ahorro en los costos de hasta el 11% respecto de un cluster HPC "tradicional".

Palabras Clave: cluster, hpc, modular, autocontenido, técnicas de construcción, cómputo paralelo.

1. INTRODUCCIÓN

En la actualidad es casi inconcebible pensar en computo intensivo sin asociarlo directamente con un cluster de computadoras personales. Un cluster de computadoras es un conjunto de computadores independientes (llamadas nodos) funcionando como si fueran un solo equipo (con la suma de los recursos de sus elementos).

Es por esto que cualquier organización que necesite capacidades de cálculo en algún momento se ve enfrentada a la posibilidad de adquirir o construir su propio cluster de computadoras. Al menos al principio, la organización optará por la construcción de un cluster de computadores personales, esto se debe principalmente a cuestiones de oportunidad, costos y escala [sterling02].

La tarea de construir un cluster (independientemente de su tamaño), involucra mucha planificación y gran cantidad de decisiones: composición de cada nodo, cantidad de nodos, comunicación de los nodos, infraestructura [steiner01] (espacio físico, soportes, alimentación eléctrica, refrigeración, protección y seguridad), sistema operativo a emplear, aplicaciones a instalar, etc.

Uno de los aspectos antes mencionados es el que suele causar mayores dificultades ya que se lo suele subestimar en la planificación, nos referimos a la infraestructura necesaria para dar soporte al cluster [feng04].

En organizaciones pequeñas suele ser complejo conseguir un espacio físico que cuente con los requerimientos de un cluster (alimentación eléctrica estabilizada y/o ininterrumpida para cada uno de los nodos, control de temperatura, control de acceso), además es necesario alojar los nodos y los dispositivos de comunicaciones de manera ordenada, lo que hace necesario estanterías y mecanismos para ordenar los cables de cada nodo [kok01] .

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Nuestro objetivo es desarrollar, documentar y verificar un conjunto de técnicas para la construcción, instalación y operación de un cluster para cálculo intensivo tratando de minimizar los inconvenientes comúnmente asociados a este tipo de actividades.

En particular nos interesa atacar los aspectos que suelen obstaculizar más gravemente un proyecto de implementación de un cluster de computadoras, y que en muchos casos, termina disuadiendo a la organización de la idea de tener su propio cluster.

Al final de nuestro proyecto habremos generado un conjunto de técnicas fácilmente transferibles y de simple implementación que permitirán a cualquier

organización de pequeño tamaño construir un cluster de pequeñas dimensiones físicas, económico y escalable a sus necesidades futuras.

Los problemas que nos interesa explorar son:

- Estrategias de construcción que minimicen el espacio requerido y los requisitos técnicos del citado espacio.
- Conjuntos de componentes que reduzcan los costos de construcción y operación, y que sean fácilmente asequibles en nuestro país.
- Estrategias que faciliten la instalación, operación y mantenimiento del cluster.

Como se mencionó anteriormente, uno de los escollos más grandes en la construcción de un cluster es la infraestructura necesaria: estanterías, conectores de alimentación, refrigeración, control de acceso (estos dos últimos que suelen hacer necesario reservar un espacio físico exclusivamente para el cluster). Para solucionar este problema proponemos como solución la utilización de racks[1] normalizados para alojar en su interior los elementos del cluster. De esta manera el cluster se encontrará "autocontenido", es decir todos sus elementos estarán dentro del rack, para ponerlo en funcionamiento sólo será necesario conectar un único tomacorriente de alimentación y un sólo patchcord a la red de computadoras de la organización.

Aclaremos que la utilización de racks en la construcción de clusters de computadoras no es para nada nuevo, todos los clusters en la lista de las computadoras más poderosas del mundo emplean racks para alojar sus nodos[top500], pero para ello utilizan hardware especializado (y bastante costoso), como ser gabinetes rackeables[2] junto a microprocesadores, disipadores y periféricos diseñados específicamente para caber dentro de estos gabinetes.

Nuestro enfoque respecto al rack difiere principalmente en que nuestro objetivo primordial en todos los casos es emplear hardware que se encuentre comúnmente en el mercado local, de esta manera nuestros resultados podrán ser fácilmente recreados por cualquier organización que lo desee. Es por ello, que en lugar de emplear gabinetes rackeables montaremos los componentes de cada par de nodos (motherboard, dispositivos de almacenamiento y fuente de alimentación) sobre bandejas normalizada[3] de 19".

Otro inconveniente importante es el consumo de energía eléctrica (con su costo y el correspondiente calor disipado), en este sentido seguiremos dos estrategias, la primera será la utilización de memorias CompactFlash (en conjunto con adaptadores CompactFlash - IDE) en lugar de emplear discos rígidos, lectoras de CD-ROM o unidades de disquetes, y la segunda, es el uso de una sola fuente de alimentación cada dos nodos.

Desde el punto de vista de la organización de las actividades, para cumplir con los objetivos antes mencionados, nuestro proyecto se encuentra dividido en tres etapas, encontrándonos actualmente al final de la primera de ellas.

La primera etapa consiste en actividades de relevamiento de las aplicaciones habituales de un cluster, los distintos tipos de clusters, las distintas técnicas de comunicación que emplean, las herramientas de uso y programación que suelen incluir y las distintas distribuciones y herramientas de instalación disponibles para instalar/configurar clusters. En simultáneo con estas actividades se realiza un relevamiento del hardware disponible en el mercado y las distintas alternativas de montaje y construcción de clusters. La finalización de esta primera etapa tendrá como resultado la construcción de un cluster con un número parcial de nodos (3 en primera instancia y 8 al final de la etapa), el que permitirá realizar verificaciones empíricas de nuestros supuestos.

La segunda etapa dará inicio con un cluster modular autocontenido con 8 nodos y utilizando éste, podremos probar y perfeccionar las distintas técnicas y estrategias de instalación y uso del cluster relevadas en la etapa anterior. En paralelo a estas actividades, se llevará a cabo un análisis de los distintos mecanismos disponibles para evaluar la performance de clusters de computadoras, los cuales servirán en la etapa final para evaluar de manera objetiva las técnicas perfeccionadas en la segunda etapa. Además, en el transcurso de la segunda etapa, se agregarán 8 nodos adicionales con lo cual al final de ésta contaremos con un cluster de 16 nodos.

En la tercera y última etapa procederemos a evaluar la efectividad y conveniencia de las técnicas perfeccionadas en la segunda etapa y generar un documento donde se encuentre plasmado el know-how adquirido a lo largo del proyecto. Al finalizar esta etapa se encontrará operando nuestro cluster modular autocontenido con un mínimo de 16 nodos. Igualmente se espera para ese momento, haber agregado otros 8 nodos, llegando al máximo de las capacidades del rack (1 coordinador y 24 nodos esclavos).

1 Un rack es un gabinete con medidas normalizadas para que sea compatible con equipamiento de diversos fabricante.
2 Gabinetes especiales diseñados para ser colocados dentro de los racks normalizados.
3 Bandejas metálicas que están diseñadas para ser colocadas dentro del racks normalizados.

3. RESULTADOS Y CONCLUSIONES PRELIMINARES

En este punto los resultados son muy prometedores. Podemos garantizar una reducción del espacio requerido de entre 23 y 30% respecto de un cluster que emplee gabinetes tradicionales en sus nodos[4], e inclusive una reducción de entre el 25 y el 75% del espacio de rack requerido respecto de un cluster armado utilizando gabinetes rackeables (que además son entre un 50 y un 150% más costosos que los tradicionales), ya que cada gabinete normalizado tiene un alto de entre 2U. (88,9mm.) y 4U. (177,8mm.)[5], cabe recodar que nuestro cluster utiliza 3U. (133,4 mm.) para alojar cada par de nodos.

Respecto al consumo de energía, nuestros cálculos y observaciones permiten estimar una reducción de entre el 15 y el 25% respecto de un cluster tradicional (cada nodo con su gabinete, fuente de alimentación y una disquetera, disco rígido o lectora de CD-ROM). Esta reducción se debe principalmente al empleo de una fuente de alimentación cada dos nodos y a la utilización de memorias CompactFlash en lugar de disqueteras, discos rígidos o lectoras de CD-ROM.

Recordemos además que una fuente de alimentación no entrega toda la potencia que consume, parte de ella (en fuentes de buena calidad no suele ser inferior al 30%) se pierde[aebischer02], por lo que al usar 12 fuentes de alimentación en lugar de 24 logramos una reducción de las citadas pérdidas. Mediciones de consumo energético realizadas en nuestro cluster experimental dieron como resultado que cada uno de nuestros nodos, tiene un consumo máximo de 71 Watts, por lo que si utilizáramos una sola fuente por cada nodo, ésta se encontraría tan subutilizada

Una memoria CompactFlash (con un adaptador CompactFlash - IDE) consume un 90% menos que una disquetera, un 95% menos que un disco rígido y 97% menos que una lectora de CD-ROM. Por lo que al utilizar éstas últimas en nuestros nodos logramos un ahorro en consumo de energía en dispositivos de almacenamiento del 91% respecto de un cluster donde todos los nodos tengan su propio disco rígido, del 94% respecto de un cluster donde los 24 nodos esclavos tengan lectoras de CD-ROM, y del 84% respecto a un cluster donde los 24 nodos esclavos usen disqueteras.

Respecto a costos de adquisición, nuestro Cluster Modular Autocontenido sería un 11% menos costoso que un cluster de 24 gabinetes independientes cada una con su disco rígido, un 2,9% más caro que un cluster con 23 gabinetes independientes con lectoras de CD-ROM además de un coordinador con disco rígido, y un 4,4% más costoso que un cluster compuesto de 23 gabinetes independientes con disqueteras de 3½" y un coordinador con disco rígido.

Cabe resaltar las ventajas de emplear memorias CompactFlash en los nodos esclavos en lugar de discos rígidos (alto costo y mayor consumo de energía), lectoras de CD-ROM (sólo lectura y alto consumo eléctrico) y disqueteras (poca capacidad, baja velocidad y pobre confiabilidad).

Claro está, la mayor ventaja de nuestro enfoque es el hecho de que nuestro cluster es modular (se le podrían anexar otros racks (cada uno con 24/26 nodos adicionales) y además está autocontenido, es decir todos componentes, cables, dispositivos de conectividad, etc. están dentro del rack, por lo cual para hacerlo funcionar sólo es necesario conectar el tomacorriente a la red eléctrica y el nodo coordinador a la red de la organización.

Otra ventaja de emplear un rack es la facilidad con la cual se podría cambiar de ubicación el cluster, e incluso, en nuestro caso, ser llevado al aula (el rack del prototipo cuenta con ruedas).

4 El cálculo se realizó teniendo en cuenta gabinetes mid-tower y sin tener en cuenta el espacio entre gabinetes, ni las estanterías necesarias.

5 Existen gabinetes con una altura inferior a 2U. pero sólo pueden alojar hardware especializado, por lo cual no fueron tenidos en cuenta.

4. REFERENCIAS

[aebischer02] B. Aebischer and A. Huser. *"Energy efficiency of computer power supply units - Final report"* - Swiss Federal Office of Energy, Technical Report 2002. http://www.bfe.admin.ch/dokumentation/energieforschung/index.html?lang=en&publication=9436

[feng04] W. Feng, Green Destiny: *"A 240 Node Energy-Efficient Supercomputer in Five Square Feet"*, http://sss.lanl.gov/presentations/041015-IEEE-DVP.pdf

[kok01] Kok, J., Elzinga, E., and Wolffe, G. *"Notes on constructing a parallel computing platform"*. J. Comput. Small Coll. 17, 1 (Oct. 2001), 71-80.

[steiner01] Steiner, S. 2001. *"Building and installing a Beowulf cluster"*. J. Comput. Small Coll. 17, 2 (2001), 78-87.

[sterling02] Sterling, T. (ed), *"Beowulf Cluster Computing with Linux"*, MIT Press, Cambridge, Massachusetts, 2002.

[top500] *TOP500*: http://www.top500.org/