Desarrollo e implementación de modelos de súper computación distribuida de bajo costo

Investigador Principal: Carlos Guillermo Bran, cbran@udb.edu.sv
Director del Centro de Tecnologías de la Información y las Comunicaciones
Vicerrectoría de Ciencia y Tenología

A través de este proyecto se diseñó un modelo escalable y eficiente de reciclado de computadoras para ser usadas en infraestructuras de procesamiento paralelo de alto rendimiento o de alta disponibilidad, usando estrategias dedicadas como los Conglomerados (Clústers) o no dedicadas como las Grillas (Grid). Esto permite la utilización de los lotes de computadoras descontinuadas pero completamente funcionales operen de forma colectiva con un sistema operativo que permita el balancear la carga de cálculo que demanda cualquier tipo de servicios en los múltiples procesadores del conjunto de estaciones para crear supercomputadoras de procesamiento paralelo que pueden ser usados en aplicaciones de alto rendimiento como el desarrollo y modelaje de sistemas de ecuaciones complejos o para el uso en aplicaciones de alta disponibilidad, lo que permite que instituciones educativas, y empresas puedan usar sus recursos tecnológicos obsoletos para sostener sus servicios y/o aplicaciones de alta demanda.

La interconexión de las estaciones hace uso de las interfaces de conexión del alta velocidad como Ethernet para la interconexión de cada una de las estaciones que forman cada celda del panal, donde se ejecutan las tareas distribuidas demandadas por cada uno de los servicios y/o aplicaciones que se ejecutan sobre toda la infraestructura.

El proyecto hace uso de tecnologías de sistemas operativos de código abierto para reducir los costos; además, plantea un conjunto de aplicaciones de ejemplo de uso, las cuales pueden dar origen a nuevos proyectos de investigación que podrían hacer uso de esta plataforma como núcleo de cálculo o arquitectura de soporte de los servicios y/o aplicaciones que demande dicho proyecto.

Antecedentes

La resolución de problemas mediante procesamiento paralelo no es nueva. Está basada en el viejo y conocido método de divide y vencerás utilizado para resolver problemas de carácter computacional.

Relativo al mundo de la tecnología y al campo de los procesadores en general, se descubrió que las arquitecturas paralelas podían solventar de manera más rápida cierto tipo de problemas. Desde 1955 personas como Gene Amdahl han investigado en el campo de arquitecturas paralelas obteniendo aquellos parámetros que optimizaban las arquitecturas así como aquellos que hacían que la relación coste-rendimiento aumentase. Empresas como IBM, DEC y desde luego muchas otras organizaciones como el MIT, se han interesado en la computación paralela desde las décadas de los 50-60, y de hecho siguen investigando y obteniendo resultados en la actualidad, hasta el punto de que prácticamente todos las computadoras que existen actualmente en el mercado explotan de una u otra manera soluciones paralelas.

En la década de los 80s, el compartimiento de recursos mediante redes de computadores hizo posible un nuevo planteamiento para aprovechar no solo recursos como capacidad de almacenamiento o

capacidad de impresión, sino para utilizar ciclos de CPU de otras máquinas conectadas a la red (los llamados multicomputadores). En los 70s y a primeros de los 80s, personas como Bruce J. Nelson de Xerox expusieron trabajos teóricos de cómo se podía utilizar mediante software esta capacidad de procesamiento paralelo. En 1985, Intel produjo el primer iPSC/1. Este multicomputador era una combinación de muchos 80286 conectados en una topología hipercubo a través de controladoras ethernet, mostrando que era real y posible utilizar este tipo de redes para explotar los sistemas paralelos. En la década de los 90, el uso de las redes de computadores se extendió de manera exagerada en comparación a otros campos como el de sistemas operativos o el de arquitectura de computadores, lo que volvió comparables las velocidades de comunicación entre nodos con las velocidades de comunicación entre chips y aceleró la capacidad de producir algoritmos para balancear la carga de un proceso entre múltiples estaciones.

Se pueden distinguir dos épocas en las cuales los problemas que han provocado la aparición de sistemas paralelos y distribuidos han sido diferentes:

Servidores de núcleo de las tres arquitecturas

- Por un lado las décadas de los 60-70-80, en las cuales el máximo problema era optimizar la capacidad de procesamiento, y de esta manera aumentar el rendimiento de las máquinas y la producción de éstas.

- Por otro lado, desde la década de los 90 hasta la actualidad, donde los problemas han aumentado. A los que existían en las décadas anteriores se han sumado los provocados por la red Internet y el fenómeno de la nueva economía, lo que ha generado una demanda creciente de poder de procesamiento para poder sostener este modelo.

Este último punto es fácil de entender: la nueva economía está formada por comercios a imagen y semejanza de los de la tradicional, pero con las ventajas aportadas por el mundo de las máquinas. Son nuevas tiendas y negocios que funcionan 24 horas al día 7 días a la semana, que no necesitan de personal, excepto técnico, para su puesta en marcha y al que se accede a través de Internet. Con este nuevo tipo de negocio, muchas empresas hacen inversiones en equipo y personal técnico, para ofrecer a nivel mundial soluciones que de otra manera podrían ser inviables por precio, tiempo u organización. Las empresas exigen a esta nuevas tecnologías, lo mismo que han exigido siempre a las antiguas:

- Máximo rendimiento, mínimo coste. Intentando hacer lo imposible para que las inversiones realizadas sean amortizadas sin desperdiciar ningún recurso.

- Máximo aprovechamiento de los recursos existentes.

- Disponibilidad máxima, lo que implica que una operación 7x24 para poder sostenerse requiere de redundancia.

- Confiabilidad máxima. Sabiendo que el sistema se va a comportar de la manera que se espera de él.

- Adaptación a los cambios. Tanto en forma de carga para el sistema como en forma de nuevo planteamiento del negocio. El sistema debe ser flexible y escalable.

La escalabilidad de un sistema es importante por motivos claramente económicos (no solo a nivel de empresa) y supone un gran reto en el diseño de sistemas para que estos puedan adaptarse de manera eficiente a nuevas exigencias. La definición de escalabilidad más apropiada a los términos relacionados con la computación distribuida es: un sistema se dice escalable si es capaz de escalar,

es decir, de incrementar sus recursos y rendimiento a las necesidades solicitadas de manera efectiva y reducir costos. Aunque la mayoría de las veces se habla de escalar hacia arriba, es decir de hacer el sistema más grande, no es siempre necesario. Muchas veces interesa hacer el sistema más pequeño pudiendo reutilizar los componentes excluidos.

Nodos de trabajo

Respecto a la evolución de los sistemas y con objeto de obtener mayor capacidad de procesamiento, el paralelismo a todos los niveles ha sido una de las soluciones más utilizadas; de hecho, en la actualidad, la totalidad de los computadoras y microprocesadores explotan, de una manera u otra, tecnologías paralelas, ya sea en multiprocesadores, en multicomputadores o en procesadores independientes por ejemplo: MMX en los procesadores Intel, 3DNow! en los AMD, Altivec en la arquitectura PPC, entre otras.

Por supuesto muchas de las soluciones de computación paralela existentes son de carácter propietario lo que contar con este tipo de recursos implica altos costos para las empresas o universidades y por otro lado las soluciones no propietarias se vuelven muy difíciles de implementar y de sostener ya que se requiere del diseño de estrategias de parametrización para acercar el problema de una institución a una aplicación ejecutable en una arquitectura paralela. La razón de esta dificultas es que no todo problema es paralelizable (ejecutarse en ambientes paralelos).

Planteamiento del problema

El problema principal a resolver es cómo lograr que nodos no homogéneos (distintas arquitecturas, capacidades y marcas), puedan operar de forma distribuida para resolver problemas o sostener servicios, manteniendo de forma optima las siguientes características:

1) **Economía:** Una relación precio-rendimiento mayor que en los sistemas centralizados o dedicados.

2) **Velocidad:** Al hacer uso de múltiples nodos, los tiempos de respuesta serán superiores a los sistemas centralizados.

3) **Alta disponibilidad:** El sistema deberá de seguir funcionando aunque se pierdan nodos del grupo.

4) **Escalabilidad:** Capacidad de crecer el poder del grupo al agregar más nodos que resultan de maquinas cuya vida útil ya ha caducado y por lo tanto no presentan mayor valor comercial.

5) **Transparencia:** Facilidad para hacer paralelizables los problemas a ejecutar en la arquitectura distribuida.

Esta última característica es la que plantea el reto real de la investigación ya que la complejidad de lograr que aplicaciones se ejecuten en ambientes paralelos de forma sencilla para el usuario final es mucho más retador y demandante que montar una arquitectura paralela en si.

Son muchas las soluciones de ambientes distribuidos (clúster) que existen tanto de forma comercial como libre, sin embargo todos estos se reducen a dos tipos principales: Clústers de alta disponibilidad (HA) y Clúster de alto rendimiento (HP).

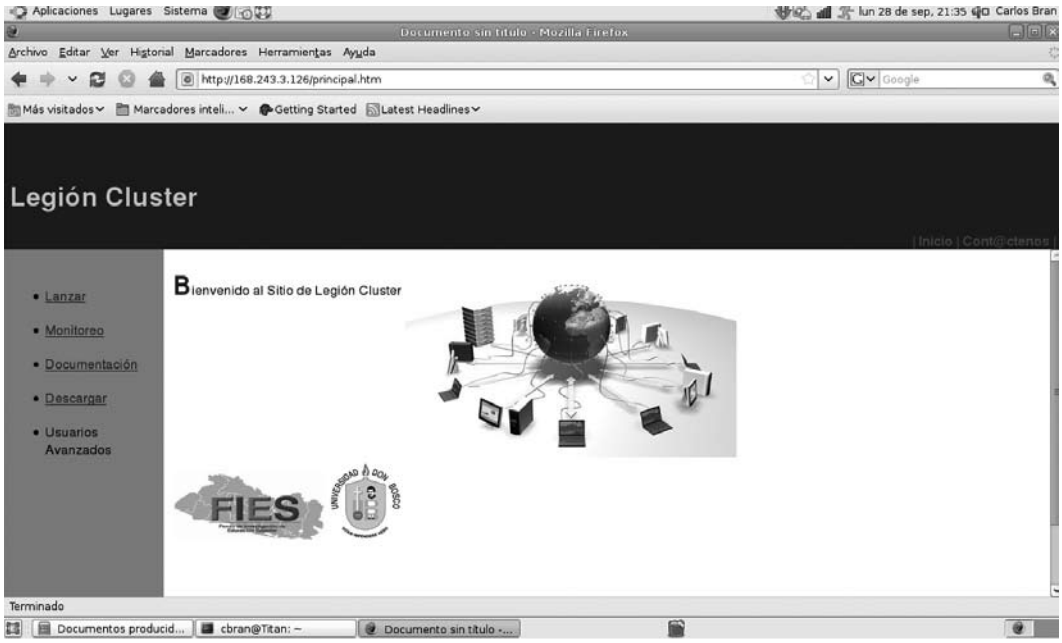
Los clústers HA están diseñados especialmente para dar un servicio de alta disponibilidad. Esto tiene muchas aplicaciones en el mundo actual donde existe gran cantidad de servicios informáticos que deben funcionar 24 horas, 7 días a la semana, 365 días al año. Estos clústers son una alternativa real a otros sistemas usados tradicionalmente para estas tareas de hardware redundante que son mucho más caros. Este tipo de soluciones pueden aplicarse para sostener servicios como los siguientes:

- Bases de datos
- Correo electrónico
- Web
- Transferencia de archivos
- Tele trabajadores

Los clúster HP están pensados para mejorar el rendimiento al procesar problemas complejos con los que se reduce el tiempo de respuesta de aplicaciones de alta demanda de procesador como las que demandan problemas tales como:

- Cálculos matemáticos complejos de ecuaciones de múltiples variables no homogéneos.
- Renderizaciones de gráficos y animación computarizada.
- Compilación de programas.
- Compresión de datos
- Descifrado de códigos
- Simulación y sistemas predictivos
- Sistemas operativos

La meta del trabajo es probar, modificar y adaptar tecnologías existentes de código abierto para soluciones de súper computación, acercando este tipo de tecnologías para que sean implementadas y utilizadas de manera fácil y rápida por empresas y universidades para montar soportar sus aplicaciones y/o servicios.



Lanzador de trabajos a la arquitectura de supercomputacion