

Public-Resource Computing: Un nuevo paradigma para la computación y la ciencia

Juan David Osorio Betancur

Psidium - Mobile Marketing
juandos@gmail.com.

Fecha de recepción: 30-06-2005

Fecha de aceptación: 01-11-2005

ABSTRACT

This article explores the Public-Resource Computing concept, an idea that has been successfully developed in the scientific community a few years ago, which involves the use of spare computational resources available in the millions of PCs in the world connected to Internet. We discuss the SETI@home project, the most successful example of this concept, and we describe the BOINC (Berkeley Open Infrastructure for Network Computing) platform, a software system created to make it easy for scientist to create and operate public-resource computing projects.

KEY WORDS

Supercomputing, Public-Resource Computing, Global Computing, Peer-to-peer Systems, Distributed Systems, SETI@home, BOINC

RESUMEN

En este artículo se explora el concepto de Computación de Recursos Públicos (Public-Resource Computing), una idea que se ha venido desarrollando con gran éxito desde hace algunos años en la comunidad científica y que consiste en el aprovechamiento de los recursos de computación que se encuentran disponibles en los millones de PC que existen en el mundo conectados a internet. Se dis-

cute el proyecto SETI@home, el más exitoso representante de este concepto, y se describe la plataforma BOINC (Berkeley Open Infraestructure for Network Computing), un sistema de software que busca hacer más fácil a los científicos la creación y mantenimiento de proyectos basados en computación de recursos públicos.

PALABRAS CLAVE

Supercomputadores, computación de recursos públicos, computación global, sistemas peer-to-peer, sistemas distribuidos, SETI@home, BOINC

Clasificación Colciencias: C

I. INTRODUCCIÓN

Blue Gene/L es un ambicioso proyecto de IBM que busca crear el supercomputador más poderoso del mundo. Su más reciente sistema beta se encuentra actualmente en la cima de la lista del TOP500 de supercomputadores con sus 70.72 TeraFLOPS de poder de procesamiento sostenido.⁽¹⁾ Sin embargo, la marca de 70 TeraFLOPS fue superada un año antes por un proyecto nacido en el seno de la academia, con un presupuesto cientos de veces menor (y obtenido por medio de donaciones), y una motivación más idealista (pero que podría partir en dos la historia de la humanidad): la búsqueda de inteligencia extraterrestre.

I.1 SETI@home

SETI@home⁽²⁾ surge como respuesta a dos de los principales obstáculos que enfrenta toda iniciativa de SETI:¹

- El afán de contar con enormes cantidades de poder de procesamiento, debido a que es necesario buscar señales débiles con parámetros desconocidos (pues nadie sabe la forma en que transmitirá una inteligencia extraterrestre), separándolas del ruido de fondo y filtrando las interferencias de origen humano provenientes de miles de fuentes conocidas y desconocidas. La premisa básica es que entre más poder de cómputo se tenga, puede hacerse

una búsqueda más amplia y con mayor sensibilidad.^(3,4)

- Los recursos económicos con los que puede contar una iniciativa de SETI son escasos en el mejor de los casos, y la mayor parte del tiempo, inexistentes, debido a que es aún un área polémica de la investigación científica. El último proyecto de SETI patrocinado gubernamentalmente (por el gobierno de EE.UU.) fue cancelado en 1992, y desde entonces el apoyo a este tipo de iniciativas se ha limitado a patrocinios secundarios de instituciones académicas, empresas e individuos. Por eso la idea de tener un supercomputador dedicado exclusivamente a la tarea es prácticamente impensable.

La idea de SETI@home es buscar el poder de procesamiento allí donde se encuentra disponible en grandes cantidades: cientos de millones de PC en todo el mundo, cuya velocidad de procesamiento crece exponencialmente por la Ley de Moore, que durante partes subsanciales de su tiempo de funcionamiento se encuentran sin uso, y que están conectados por una red: internet.

Aun cuando SETI@home no fue el primer proyecto² concebido para explotar estos 'recursos públicos', se convirtió rápidamente en el más exitoso, atrayendo a millones de usuarios dispuestos a donar el tiempo li-

1. Search of ExtraTerrestrial Intelligence (Búsqueda de Inteligencia Extraterrestre): Área de la ciencia cuyo objetivo es detectar vida inteligente por fuera de la Tierra. En este artículo se habla de SETI refiriéndose a radio SETI, su más importante rama, que emplea radiotelescopios para detectar señales de radio de banda estrecha (que no ocurren naturalmente) provenientes del espacio exterior.³

2. En 1997, dos años antes de su lanzamiento, surgieron dos proyectos que atrajeron a miles de participantes: GIMPS, que buscaba números primos extremadamente grandes (números de Mersenne), y Distributed.net, cuyo objetivo era romper por fuerza bruta algunos protocolos de encriptación.

bre de sus PC por medio de un salvapantallas que proporciona una vista gráfica del procesamiento de señales en proceso. Desde su lanzamiento en 1999, más de cinco millones de usuarios se han registrado en el proyecto y un año después ya superaba las marcas de los supercomputadores del TOP500. A la fecha, más de un millón de usuarios continúan activos, proporcionando más de 70 TeraFLOPS de poder de procesamiento sostenido. Además, el proyecto ostenta la marca de ser la computación más larga jamás realizada (1.5 millones de años de CPU, a marzo de 2004).

El éxito de SETI@home inspiró a otros proyectos académicos con necesidades similares. Sin embargo, hasta hace poco un número relativamente pequeño de este tipo de proyectos había tenido éxito, debido en parte a la falta de un apropiado middleware (software de cliente y de servidor, herramientas de administración, características del lado del usuario, etc.).

2. CONQUISTANDO EL PARADIGMA DE LA COMPUTACIÓN DE RECURSOS PÚBLICOS (PUBLIC-RESOURCE COMPUTING)

La computación de recursos públicos (llamada también computación global y computación peer-to-peer) ofrece recursos sin precedentes para la investigación científica y la ingeniería. Se espera que para el 2015 haya mil millones de computadores personales

conectados a internet; el poder de procesamiento que representarán es del orden de varios PetaFLOPS. Más aún, la tasa de crecimiento de la velocidad de los coprocesadores gráficos, usados en las tarjetas aceleradoras de vídeo en PC y en consolas de juegos, es incluso mayor que la de los procesadores (se duplica cada ocho meses), y su poder de procesamiento de punto flotante puede ser varias veces mayor que el de una CPU, con una enorme y hasta ahora no explotada fuente de recursos de cómputo.³ Además, la computación de recursos públicos posee otros aspectos. En cuanto al almacenamiento, por ejemplo, si 100 millones de usuarios de PC donan cada uno 10 GB de almacenamiento, se alcanza una cifra de 1 Exabyte (10^{18} bytes), que excede con creces la capacidad de cualquier sistema de almacenamiento centralizado.⁽⁶⁾

Sin embargo, para poder aprovechar esta abundancia de recursos es necesario superar una larga serie de retos técnicos: Adaptar una aplicación a una multitud de plataformas, implementar sistemas de servidor y bases de datos, mantener el registro de las cuentas de usuario y sus respectivos créditos (contribución de cada usuario al proyecto), manejar la redundancia y las condiciones de error, además de prever acciones malintencionadas de usuarios, y un largo etcétera.

2.1 BOINC

La infraestructura abierta para computación en red de Berkeley (BOINC,

3. SETI@home está trabajando con NVIDIA para la creación de una versión de la aplicación que realice una gran parte del procesamiento en el procesador de las tarjetas de video de éste fabricante.

por sus siglas en inglés: Berkeley Open Infrastructure for Network Computing^(7,8)) fue desarrollada por el Laboratorio de Ciencias del Espacio de la Universidad de California - Berkeley, con el fin de “apoyar el desarrollo del paradigma de la computación de recursos públicos”, por un lado facilitando a la comunidad científica la creación de proyectos de este tipo, y por el otro, alentando al público en general a participar en proyectos, al brindarles una forma simple de unirse a una o varias iniciativas, y controlar el uso que éstas hagan de sus PC.

Los objetivos específicos de BOINC incluyen:

- *Reducir las barreras de entrada a la computación de recursos públicos.* BOINC está diseñado para ser usado por científicos, no por ingenieros de sistemas o expertos en programación. Las herramientas que provee son simples y bien documentadas, de manera que un proyecto pueda ser creado con relativamente poco trabajo inicial, y pocas horas semanales de mantenimiento. Además, el servidor central de un proyecto basado en BOINC puede consistir en una sola máquina que ejecute software de código abierto de uso común, tales como Linux, Apache, PHP, Python y MySQL.
- *Compartir recursos entre proyectos autónomos.* Aun cuando los proyectos basados en BOINC son autónomos (operan en sus propios

servidores, y corren completamente independientes unos de otros), un usuario puede participar simultáneamente en múltiples proyectos, definiendo para cada uno qué porcentaje de los recursos de su máquina puede utilizar. De hecho, cuando un usuario se registra en varios proyectos la utilización de sus recursos mejora, puesto que así si un proyecto se encuentra abajo por mantenimiento, el PC del participante puede utilizar sus recursos en otros proyectos, en lugar de simplemente esperar a que se restablezca la comunicación con el primero.

- *Soportar aplicaciones diversas.* BOINC soporta un amplio rango de aplicaciones, incluyendo las ya existentes y escritas en lenguajes comunes (C, C++, FORTRAN), sin necesidad de realizar modificaciones extensas.
- *Recompensar a los participantes.* Para que un proyecto de computación de recursos públicos sea exitoso, debe atraer y retener participantes, y para lograrlo debe ofrecerles incentivos. El más importante para la mayor parte de los participantes⁴ es el crédito: una medida numérica de los recursos que ha brindado al proyecto. BOINC provee un sistema de créditos por participante y por proyecto, que es altamente resistente a fraudes por parte de usuarios malintencionados. Además, BOINC facilita a los desarrolla-

4. Según encuestas realizadas a participantes de SETI@home.

dores de un proyecto adicionar gráficos de salvapantallas a sus aplicaciones, aspecto importante dado que éstos son otros de los incentivos que más importan a los participantes.

2.2 La Implementación de BOINC

Un proyecto basado en BOINC se identifica con una URL maestra, que corresponde a la página principal del sitio web del proyecto, y funge como directorio de los servidores del mismo. Un proyecto involucra una o más aplicaciones, que pueden cambiar en el tiempo en forma transparente para el usuario.

El complejo de servidores de un proyecto basado en BOINC está centrado en una base de datos relacional que almacena información acerca de las aplicaciones, plataformas, versiones, unidades de trabajo, resultados, cuentas de usuarios, equipos de usuarios, etc. Las funciones de servidor son realizadas por un conjunto de servicios web y procesos demonio:

- Los Servidores de Programación de Trabajo (Scheduling Servers) manejan las RPC de los clientes, que pueden consistir en solicitudes de trabajo, envío de resultados, o solicitudes de actualización de datos.
- Los Servidores de Datos (Data Servers) manejan los envíos de resultados por parte de los clientes, realizando un proceso de verificación de los mismos para comprobar su validez. Tanto el envío de trabajo a los clientes como la recepción de resultados de los mismos se manejan por HTTP, de manera que los clientes pueden funcionar detrás de un firewall.

BOINC provee herramientas (scripts de Python e interfaces de C++) para crear, iniciar, y detener proyectos, y comunicarse con ellos; adicionar nuevas aplicaciones, plataformas o versiones a un proyecto; crear unidades de trabajo y monitorear el desempeño de los servidores y de los clientes.

Para participar en proyectos basados en BOINC, un usuario debe descargar el cliente de BOINC. Este puede operar en diversas modalidades:

- Como salvapantallas, mostrando gráficas asociadas al proyecto (que es el modo más usual).
- Como servicio de Windows, que puede correr aun si un usuario no ha iniciado sesión en el equipo; como una aplicación que muestra información acerca del proyecto y el uso del equipo.
- Como un programa de línea de comandos de UNIX.

BOINC también provee herramientas para realizar la instalación remota del cliente en múltiples máquinas.

Un participante se une a un proyecto visitando el sitio web del mismo, llenando la forma de inscripción y registrando el proyecto en el cliente. Para esto debe escribir la URL maestra del proyecto y un código de participante que le es asignado una vez realizada su inscripción. El cliente de BOINC se ocupa de descargar las aplicaciones y archivos necesarios, sin más participación del usuario

3. APLICACIONES DE LA COMPUTACIÓN DE RECURSOS PÚBLICOS

Como se ha visto, la computación de recursos públicos ofrece posibilidades

sin precedentes para diversos campos de la ciencia y la ingeniería, y ya existe un importante marco de trabajo para la creación de proyectos que exploten este paradigma. Sin embargo, dadas las características propias de la computación de recursos públicos, se presentan ciertas limitaciones al rango de iniciativas que pueden aprovechar este concepto:

- Las tareas deben tener una alta relación de cantidad de cómputo a cantidad de datos, es decir, se debe poder realizar una gran cantidad de computación sobre un grupo relativamente pequeño de datos (Por ejemplo, SETI@home realiza cerca de doce horas de computación sobre cada unidad de 350 KB de datos, produciendo resultados de 1 KB). Esto con el fin de mantener bajo el tráfico de red, y por consiguiente el costo de conectividad del proyecto.
- Las tareas deben ser inherentemente independientes. Existen aplicaciones que pueden ser fácilmente trabajadas en paralelo, pero que requieren una frecuente sincronización entre nodos. Este tipo de tareas no son compatibles con las frecuentes fallas de comunicación que se darán entre los participantes y el servidor central de un proyecto de computación de recursos públicos.
- Las tareas deben ser tolerantes a errores. Los participantes de un proyecto pueden generar resultados erróneos al surgir fallas en el sistema operativo o en el hardware de sus PC, o bien en el cliente BOINC en caso de que lo hayan modificado y recompilado para tratar de obtener un mayor ren-

dimiento; o generar resultados deliberadamente falsos para tratar de obtener un mayor crédito, o simplemente para vandalizar el proyecto.

- Y quizás lo más importante, un proyecto de computación de recursos públicos debe atraer participantes.

Esto último añade un importante componente al paradigma: el aspecto social. Los creadores de un proyecto tienen que hacer mucho énfasis en la labor de atraer y mantener participantes, y para esto se han identificado diversas estrategias. En primer lugar, las metas del proyecto deben tener una alta aceptación (Por ejemplo, existen iniciativas exitosas en campos como la investigación biomédica y el estudio del calentamiento global). Además, éstas deben ser explicadas y justificadas con claridad. En segundo lugar, las gráficas de salvapantallas pueden ser en extremo importantes para algunos participantes, y constituyen un poderoso mecanismo promocional puesto que pueden ser vistas por muchas personas que podrían interesarse en participar. En tercer lugar, debe crearse una comunidad alrededor de un proyecto, donde los usuarios puedan compartir opiniones, consultar su crédito y compararlo con el de otros participantes, crear perfiles de usuario y formar equipos. El factor motivacional de pertenecer a una comunidad que apoya directamente un proyecto con metas importantes no se debe descuidar.

CONCLUSIÓN

La computación de recursos públicos ofrece posibilidades sin precedentes a la ciencia y la ingeniería. El tener a

disposición potencia de cómputo y almacenamiento virtualmente ilimitados permitirá la creación y desarrollo de proyectos de investigación con niveles de complejidad que hubieran sido imposibles unos años atrás. El nivel de precisión con el cual los modelos matemáticos se aproximan al mundo físico, y el detalle con el que pueden analizarse enormes cantidades de datos, pueden crecer en varios niveles de magnitud mediante el uso de este potencial. Sin embargo, esto plantea un nuevo reto a los investigadores: hacer mercadeo y publicidad adecuados de sus proyectos, aprender a conquistar al público y convencerlo para que participe.

BIBLIOGRAFÍA

1. TOP500: <http://www.top500.org/lists/2004/11/>
2. SETI@home: <http://setiweb.ssl.berkeley.edu/>
3. Anderson, D. P., J. Cobb, E. Korpela, M. Lebosfsky, and D. Werthimer. "SETI@home: An Experiment in Public-Resource Computing". *Communications of the ACM*, Nov. 2002, Vol. 45 No. 11. (<http://setiathome.berkeley.edu/cacm/cacm.html>)
4. Korpela, E., D. Werthimer, D. P. Anderson, J. Cobb, and M. Lebosfsky. "SETI@home - Massively Distributed Computing for SETI". *Computing in Science and Engineering* 3(1), 2001

5. Skawinnski, K. Z. "IT & SETI: The Role of Computer Technology in the Search of Extraterrestrial Intelligence". *California Computer News*, July 2 2002. (<http://www.ccnmag.com/story.php?id=146>)
6. Anderson, D. P. "Public Computing: Reconnecting People to Science" Conference on Shared Knowledge and the Web, Residencia de Estudiantes, Madrid, Spain, Nov. 17-19 2003. (<http://boinc.berkeley.edu/boinc2.pdf>)
7. Anderson, D. P. "BOINC: A System for Public-Resource Computing and Storage" 5th IEEE/ACM International Workshop on Grid Computing, November 8, 2004, Pittsburgh, USA. (http://boinc.berkeley.edu/grid_paper_04.pdf)
8. BOINC: <http://boinc.berkeley.edu>

CURRÍCULO

Juan David Osorio Betancur. Ingeniero Telemático de la Universidad Icesi. Ha realizado investigaciones en el área de tecnologías inalámbricas y computación distribuida vinculado al grupo I2T de la misma Universidad. Actualmente es socio fundador y gerente de tecnología de Psidium - Mobile Marketing, empresa dedicada a la prestación de servicios de valor agregado sobre telefonía celular. ☀