我们身边的分布式计算

2012年2月27日

人人都可参与分布式计算项目

分布式计算专家韦伯教授曾说：

信息的恶性增长使各个领域都布满了吞噬巨大计算力的‘黑洞’，但这并不可怕，真正可怕的是我们总是想用为数不多的几块巨石而不是取之不尽的泥土去填平它。

巨人搬不动的石头，动用成千上万只蚂蚁也许就能搬动它

人人都可参与分布式计算项目

目前分布式计算项目已经有很多：

- 天文学
- 生命科学
- 数学、密码学
- 计算机科学

科学研究的前沿领域和持久感趣的方面：

- 地外文明
- 生命起源

多数项目，只要你有兴趣，就可以参与进去。

人人都可参与分布式计算项目

几个经典的项目：

- 寻找外星人：SETI@home
- 寻找梅森素数
- Google： Majestic-12

几款应用软件：

- Napster
- BitTorrent (BT)

SETI@home简介

SETI@home是Search for Extraterrestrial Intelligence at Home的缩写，为“在家里搜索地外文明”之意。

这个项目由美国加州大学伯克利分校“搜寻地球外智能”(SETI)研究小组发起，旨在利用因特网中不计其数的计算机的闲置时间进行SETI计算，以期从海量的信号中搜寻到地外文明的蛛丝马迹。

工作原理

首先由位于波多黎哥群山之中的巨型射电望远镜Arecibo收集地外信号，然后将每天约35 GB的数据传送到SETI@home项目管理中心。SETI@home管理中心将数据进行分解处理，划分成合适的大小，然后通过因特网将它们分发到全球成千上万志愿者的电脑中。

SETI@home程序在志愿者的个人计算机上，通常在屏幕保护模式下或以后台模式运行。它利用的是多余的处理器资源，不影响用户正常使用计算机。当一个信号单元分析完毕，客户端程序将有价值的信号送回SETI@home项目管理中心并自动下载新的数据。

如果志愿者送回的处理结果经确认属重大发现，那么志愿者将同SETI @ home项目组共同分享“发现者”的荣誉。

File Settings Help

The Search for Extraterrestrial Intelligence at HOME

Press F1 for info Version 3.08
http://setiathome.berkeley.edu

Data Analysis

'Getting data - connecting to server. 100%

Doppler drift rate: 0.0665 Hz/sec Resolution: 0.149 Hz

Best Gaussian: power 1.27, fit 3.149

Overall: 100.000% done CPU time: 0 hr 10 min 57.6 sec

Data Info

From: 44 hr 37' 58" RA, + 26 deg 13' 11" Dec

Recorded on: Fri Jan 07 17:27:09 2005 GMT

Source: Arecibo Radio Observatory

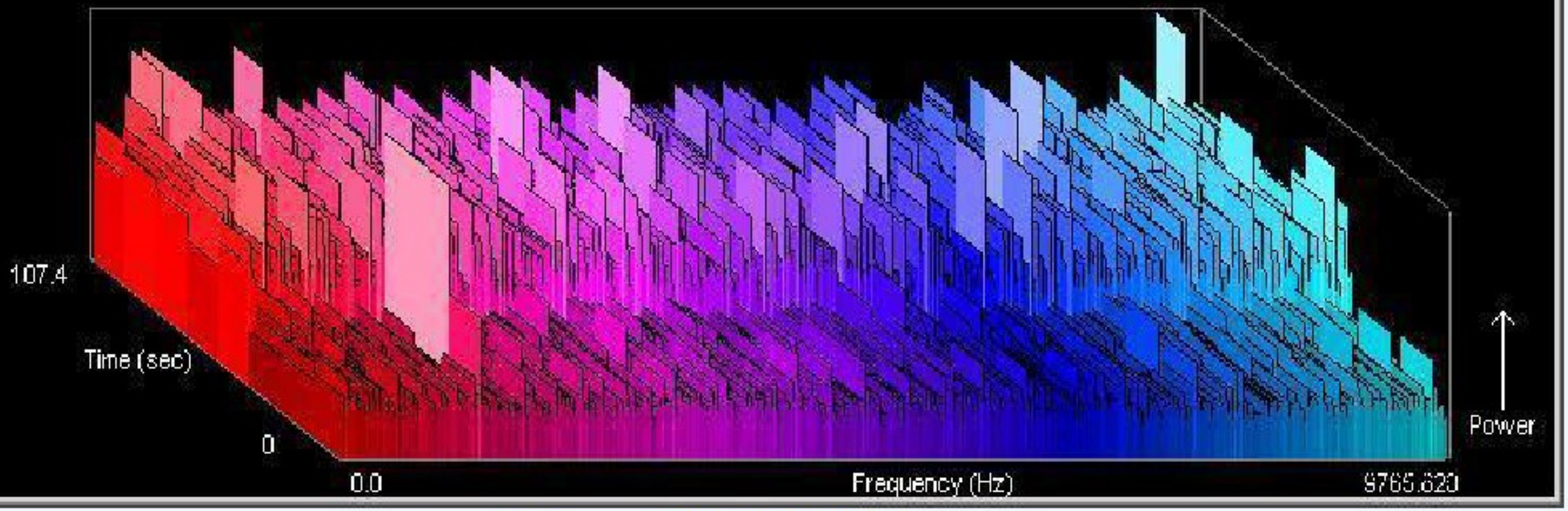
Base Frequency: 1.419736328 GHz

User Info

Name:

Data units completed: 9

Total computer time: 196 hr 13 min 20.0 sec



SETI @ home运行时的样子

成果

- SETI@home项目自1999年5月17日开始正式运行。

- 至2004年5月，该项目在世界各地拥有近500万参与者，积累了近200万年的CPU运行时间，进行了近 5×1021次浮点运算，处理了超过13亿个数据单元，无疑是非常成功的分布式计算试验项目。

- 截至 2005 年关闭之前，它已经吸引了 543 万用户，这些用户的电脑累积工作 243 万年，分析了大量积压数据，但是项目没有发现外星文明的直接证据。

寻找梅森素数，赢十万美元大奖

简介

GIMPS（Great Internet Mersenne Prime Search）因特网梅森素数大搜索，是一个数学领域的分布式计算项目。由于“电子边界基金”(Electronic Frontier Foundation)宣布将向第一个找到超过1000万位梅森素数的个人或机构颁发十万美元的奖金，使它成为为数不多的有奖金的项目，令志愿者趋之若鹜。

高阶的梅森素数具有可怕的长度，要验证一个这样的大数是否是素数，计算量大得惊人，同著名的大数质因子分解的难度有一比，而后者恰是现代公开密钥技术RSA的数学基础。经过几百年来的努力，人们才发现41个梅森素数，而其中就有7个是GIMPS项目的成果。目前最大梅林素数也是通过GIMPS项目找到的。


Prime95
Test Advanced Options Help
Sending text message to server.
UID: zzh2005zzh2005/howjoy2005, UID: zzh2005zzh2005, User: Sunny Zou, zzh2000@sina.com
Getting exponents from server
Sending expected completion date for M24921451: Jun 22 2005
Done communicating with server.
The program will now perform a self-test to make sure the
Lucas-Lehmer code is working properly on your computer.
This will take about an hour.
Test 1, 3100 Lucas-Lehmer iterations of M24903681 using 1280K FFT length.
Contacting PrimeNet Server.
Test 2, 3100 Lucas-Lehmer iterations of M24903679 using 1280K FFT length.
Test 3, 3100 Lucas-Lehmer iterations of M24892961 using 1280K FFT length.
Test 4, 3100 Lucas-Lehmer iterations of M23892959 using 1280K FFT length.
Self-test 1280K passed!
Trying 1000 iterations for exponent 24921451 using 1280K FFT.
If average roundoff error is above 0.24216, then a larger FFT will be used.
After 100 iterations average roundoff error is 0.23594.
After 200 iterations average roundoff error is 0.23782.
After 300 iterations average roundoff error is 0.23793.
After 400 iterations average roundoff error is 0.23803.
After 500 iterations average roundoff error is 0.23722.
After 600 iterations average roundoff error is 0.23714.
After 700 iterations average roundoff error is 0.23685.
After 800 iterations average roundoff error is 0.23663.
After 900 iterations average roundoff error is 0.23652.
Final average roundoff error is 0.23642, using 1280K FFT for exponent 24921451.
Starting P-1 factoring on M24921451 with B1=370000, B2=370000
Chance of finding a factor is an estimated 2.57%
P-1 on M24921451 with B1=370000, B2=370000

梅森素数计算界面

工作原理

GIMPS的工作原理与SETI@home类似，也是将庞大的数据量分成小块，再通过为数众多的客户端进行计算。

● GIMPS客户端程序可在网上下载。在首次运行时，需要输入用户ID、电脑ID，并对CPU占用率、内存占用率以及开放的计算时间等选项进行设置，一般选默认值即可。

●之后程序开始对电脑进行测试，以确定该电脑是否适合参加GIMPS项目。这个测试费时颇多，需要参加者有些耐心。

●测试完成后，便开始从服务下载数据片断进行计算。一个片断计算完后，GIMPS客户端程序会自动到服务器上下载新片断。

如果你的运气实在好，新的梅森素数恰在你计算的片断内，那么你不仅能得一大笔钱，而且还能青史留名。当然，中奖概率肯定比摸中500万体彩大奖还要低得多，所以也不必特别在意是否可以赢取奖金，权当了回国际义工。

打造分布式Google: Majestic-12

搜索成就了因特网的老大Google，也吸引了许多公司对搜索的狂热追捧。作为未来因特网世界的重要构建者，分布式计算没有理由不染指其中，何况从原理上说，分布式搜索引擎比现在的各种搜索引擎更为强大，因为它可有无数个信息“钻探机”。

Majestic-12就是这样一个基于分布式计算原理的因特网搜索引擎研究项目(Distributed Search Engine Project)，它在客户端使用一种名为“crawls”(爬行者)的技术来监视指定的网站，以便及时了解这些站点内容的变化情况，以便随时更新存放于项目服务器上的查寻索引文件。目前，该项目已对10亿 多个 URL 地址建立了完善的索引，其搜索容量已直逼一些著名的搜索引擎。

Majestic-12客户端程序

如果你乐意在搜索上做点事情的话，不妨参加这个研究性质的项目。它不仅可以使你了解被某些厂商宣传得有些神秘的搜索内幕，而且你还可以分享到在许多方面并不亚于Google的搜索结果，而这些结果也许恰是你的计算机搜罗和整理的。

近日，Majestic-12项目组推出了可利用Majestic-12成果的Firefox 搜索插件，看来可能会对Google形成威胁的新一代搜索引擎就要从地下冒出来了。

Napster 简介

这是一款可以在网络中下载自己想要的MP3文件的软件名称.它同时能够让自己的机器也成为一台服务器，为其它用户提供下载。在这个网络中，Napster本身并不提供MP3文件的下载，它实际上提供的是整个Napster网络的MP3文件“目录”，而MP3文件分布在网络中的每一台机器中，随时供你选择取用，我们下载都是直接连到另外一台机器。传输速度也相当惊人。

Napster具有强大的搜索功能，可以将在线用户的MP3音乐信息进行自动搜寻并分类整理，以备其他用户查询，只要知道你喜欢歌曲的名称或演唱者的名称，就可以和全世界乐迷共享丰盛的音乐大餐。你可以选择自己要与其他人在网上共享的音乐文件的目录，并且可以与喜欢同样风格音乐的人聊天、在论坛讨论，互相交流。

什么是分布式计算(Distributed Computing)呢?

分布式计算是计算机科学的一个重要分支，主要研究如何把一个需要巨大的计算能力才能解决的问题分解成许多小的部分，然后把这些部分分配给许多计算机进行处理，最后把这些计算结果综合起来得到最终的结果。这是一个比较狭窄的定义。一般认为，凡是基于分布式计算原理的所有应用，都应归于分布式计算的范畴，包括许多完全或部分摆脱了客户/服务器模式的新型网络软件，尤其是当下十分流行的P2P文件交换软件。

提示:P2P是一种不依赖服务器的通讯方式，与网络分布式计算如影随形，然而它并非分布式计算的要件，也就是说，使用了P2P技术的软件并非都属于分布式计算范畴。只有那些主要计算工作在客户端完成而仅使用P2P作为通讯手段的软件，才可以归到网络分布式计算程序类中。

分布式计算已在我们身边

Napster和BT都是典型的网络分布式计算程序。如果从广义的分布式计算的定义来看，我们经常使用的QQ、MSN等即时通讯工具，虽然采用的是传统的客户/服务器架构，但它在音频视频的播放和文件的传输上，使用的却是P2P技术，因此仅就此点而论，QQ等也算半个分布式计算程序。

分布式计算的发展史

1993年，DEC系统研究中心的研究员Lenstra和Manasse召集了600名志愿者，利用分布式计算方法参与由著名的美国RSA研究所发起的RSA-129密码破译活动，并在很短的时间内成功破译密钥。这次活动使人们见识到分布式计算的威力，此后对它的研究空前活跃起来。

1995年，分布式计算再接再厉，一举攻破了RSA-130。这是一个130位加密算法，这次活动开启了分布式计算和因特网结合的大门，使分布式网络计算成为主流的研究方向，并最终导致网格的诞生。

1996年，著名的GIMPS(互联网梅森素数大搜索)项目开始启动。近十年来，通过它已发现多个梅森素数。

1999年，著名的寻找外星智能生命信息的SETI@home项目正式推出，它以无比的神秘感吸引了因特网上数百万名志愿者，成为目前参加人数最多的分布式计算项目。

2000年，19岁的大学生Shawn Fanning开发出Napster，在网上掀起网络音乐交换热潮，催生了一个庞大的在线音乐市场。

2001年，IBM公司宣布自己的网格研究计划，并将为此投资40亿美元。

2002年，由Bram Cohen开发的分布式下载工具BitTorrent(BT)横空出世，以革命性的面目改变了传统的网络交换方式。

2003年，IBM发起史上最大网格运算计划，共有1000万台电脑连入其中。

分布式计算的未来

如何动员和利用社会中丰富的计算能力，始终是一个充满挑战性的问题。

可以预见，纯粹的分布式计算项目将会越来越多，它将为那些需要强大的计算能力的领域提供服务，同时它也会在网络安全和军事方面得到更多应用，甚至能成为信息战的超级武器。

网络分布式计算将继续成为热点

从主要应用于音乐和软件交换扩散到视频共享、 网络电话、视频聊天、网络游戏和IPTV等方面。

网络游戏方面，网络分布式计算也将大有用武之地，甚至能一举改变现有的网络游戏运行模式。也许有一天，网络游戏再也不需要数量庞大的服务器群，玩家可以P2P方式直接交互，使因特网成为一台庞大无比的游戏机。而通用的分布式游戏平台也将诞生，它既能提高网络游戏开发的效率，减少网络游戏的运营成本，而且还能为玩家带来莫大的方便。这样一个平台，也许会集成在未来的桌面操作系统中