P

Public Grid Computing Projects Survey and Analysis

Nolan J. Taylor
Indiana University, USA

PUBLIC GRID COMPUTING

Personal computer performance has increased dramatically over the past decade. So powerful are these machines that they complete most required tasks with time to spare. This spare time is spent in an idle loop, waiting for the next task. A number of organizations have launched public grid computing projects (i.e., using Internet-enabled personal computers) to capture these "lost" CPU cycles via specialized software that effectively aggregates the power of separate systems into a virtual computer. In many respects, computing grids are analogous to electrical power grids (Robb, 2002). Plugging in an electrical power-consuming device taps into the unused capacity of an entire system of power stations and not necessarily the one closest generation station. In computing grids, the individual computers make up a "grid" of processing power stations and the PC owners the associated virtual community.

These public grids represent an important social phenomenon because of both their membership size and the problems they address. Among the largest of these virtual organizations is the SETI@home project. Launched in May, 1999, the SETI@home (Search For Extraterrestrial Intelligence) project uses a grid to analyze radio telescope data for patterns indicative of intelligent origins (Anderson, Cobb, Korpela et al., 2002). This virtual community of PC users who have "volunteered" their PCs to aid in the project's mission currently exceeds 4 million volunteers. Grids have also been formed to assist in solving a variety of problems ranging from discovering treatments for diseases such as HIV/AIDS and cancer to math-based problems such as prime number searches and high precision calculations of the constant PI (π).

One of the most common types of public grid projects focuses on drug evaluation. This process involves screening literally millions of chemical compounds to identify candidates for use in drug design. The world community grid is one of the largest projects of this type. IBM recently announced the world community grid and its stated goal of creating the world's largest public computing grid for advancing human welfare (McDougall, 2004). The grid's initial project is processing for the Human Proteome Folding Project using the data generated from the Human Genome Project to discover the functions of the proteins that are encoded in human DNA. This information will assist in the formulation of disease cures. The infrastructure is intended to be made available to process other tasks for not-for-profit organizations in the future.

Public Grid Computing: The Basics

Public computing can be defined as massive parallel distributed computing using the Internet for data transfer to build a centrally controlled meta-computer from the idle CPU cycles of PCs of voluntary participants from the general public (Loewe, 2002, p. 378). Unlike private grids where the assets belong to a single organization using a private network protected by a firewall, public grids (also referred to as global, community, public-resource, and free grids) weave a virtual organization of PC users using the Internet as the connecting fabric.

Along with the network infrastructure (i.e., the Internet), public computing grids have three major components: clients, participants, and a grid operator. The participants consist of individuals or organizations that are willing to accept a portion of the overall computational task on their machines. The grid operator is responsible for the overall operation of the grid, development of the client and server software, allocating tasks of the various clients among the various providers, and the overall distribution of data and collection of results. Clients are the individuals or organizations with the computational task(s) and data. In some cases, the client and grid operator may be the same organization (e.g., SETI@home).

Benefits and Challenges of Establishing a Public Computing Grid

Public grid computing offers a number of benefits (Table 1). The first group of benefits centers on its lower costs as compared with monolithic computing solutions. Networking costs are reduced by using the Internet and by using "volunteered" processing power rather than fixed assets. The second group of benefits stem from its faster results as compared with monolithic computing solutions. Because virtual computers with performance in excess of the most powerful supercomputer can be constructed in this manner, tasks can be performed in less time (Gabriel Consulting Group, 2004; IBM Corp, 2004). This is particularly important in computationally intensive tasks. For example, the entire process of leading up to the bringing a new drug to market can take as many as 15 years using traditional methods (Lunney, 2001).

Not all tasks however can be efficiently executed in a public grid environment. To be efficient, a task must be parallelizable. That is, the overall task is divisible into parts that can be performed with some degree of independence from each other on separate processors. In accordance with Amdahl's law, the greater the proportion of the overall task that is parallelizable, the greater the potential "speedup" from using parallel processing (Lewis & Hesham, 1992). In the ideal case, the so-called "embarrassingly parallel" task, the problem can be partitioned such that each subtask can be performed autonomously. In addition to being parallelizable, the time required for processing must exceed the overhead (i.e., the additional time for segmenting, transmitting, and reassembling data) by a substantial margin to be efficient. In parallel processing parlance, this ratio is referred to as granularity. The higher the ratio of processing time to communication time, the greater the performance improvements in a parallel processing environment (Golub & Ortega, 1993). Additional challenges lie in effective integration of geographically dispersed, heterogeneous resources, especially given the unpredictability of both the Internet and individual PCs (Smith & Konsynski, 2004).

RECENT APPLICATIONS OF PUBLIC GRID COMPUTING

In the last 10 years, "public" grids have been formed. Initial uses for such networks involved high precision mathematical calculations (PiHex), prime number searches (Great Internet Mersenne Prime Search), and code-breaking (The CS-Cipher Challenge). More recently, grids have been initiated to assist in solving problems ranging from discovering new prime numbers to searching for treatments for HIV/AIDS. A list of major public grid projects is summarized in Table 2. Many of these grids have become quite sizable with over a million participants. Public grids also include efforts where the processing is primarily human (requiring human intelligence and inputs) such as Distributed Proofreaders.

Table 1. A summary of public grid computing benefits (Source: Gabriel Consulting Group, 2004; IBM Corp, 2004)

Savings
- The Internet serves as the connection method reducing the need to invest in network infrastructure.
- Much of the computing resource costs as well as operating costs (e.g., upgrades, electricity, etc.) are borne by the volunteers.
- Computing resources can be expanded by simply increasing the number of volunteers with less expense than would be required to increase equivalent supercomputer capabilities.

Faster Results
- Higher performance allows tasks to be completed in a shorter period of time.
- Faster results can result in savings in terms of time as well as lives when the research is related to life-threatening issues.
- The additional processing capacity may give rise to completely new opportunities that were previously unavailable because of processing limitations (e.g., for simulation of phenomena that were previously too computationally intensive to perform)