

Distributed Computing Technologies and Their Application to Drug Discovery

Andrew A. Chien
Chief Technology Officer, Entropia, Inc.
SAIC Chair Professor, UC San Diego

Distributed Computing, the exploitation of idle cycles on pervasive desktop PC systems, offers the opportunity to increase the available computing power by orders of magnitude (10× to 1000×). Such large-scale resource sharing is a key part of the emerging “Grid” computing technologies being developed and pursued by a broad array of researchers, software vendors, and hardware vendors. However, for desktop PC distributed computing to be widely accepted within the enterprise, the systems must achieve high levels of robustness, security, scalability, unobtrusiveness, and manageability. In addition, as with any novel platform technology, the systems must also capture a critical mass of applications that make the platform valuable.

We describe the emerging distributed computing technologies, focusing in particular on their system architecture and approaches to solve the key challenges. We will describe the Entropia system as a case study, detailing its internal architecture and philosophy in attacking these key problems. In particular, key aspects of the Entropia system include the use of

- scalable web/database technology for system management
- network tunneling and application namespaces for logical connectivity
- binary sandboxing technology for security and unobtrusiveness
- open integration model to allow applications from many sources to be incorporated

We describe the Entropia system and how these technologies are combined to produce a robust, flexible, high-performance system which is in use in numerous enterprises supporting a wide range of applications.

One promising area for distributed computing is a cluster of applications that support early drug discovery. Computational demands in this area are growing in accord with Venter’s Law, a more rapid increase than Moore’s Law. We discuss applications from Bioinformatics or Computational Chemistry, which all involve large numbers of parallel runs without dependencies between them (“embarrassingly parallel” jobs). In addition, most of the applications involve the use of significant quantities of data (either sequence or molecular databases), and large amounts of computation. We will describe several of these applications and give examples of how they are used in early drug discovery—a critical factor that determines their scale-up needs. We also describe their performance in a distributed computing system.