High-Performance Artificial Intelligence

James Hendler

One of the most exciting changes in computing in the past decade has been the introduction of high-performance computing (HPC) made possible by the advent of parallel supercomputers. In the field of artificial intelligence (AI), some researchers have realized that the enormous computing power and large memory of these parallel computers open a new horizon for research. The ability to automatically create and search massive "knowledge bases" provides a key to scaling AI beyond its current limits and may provide a critical technology in the emerging national information infrastructure.

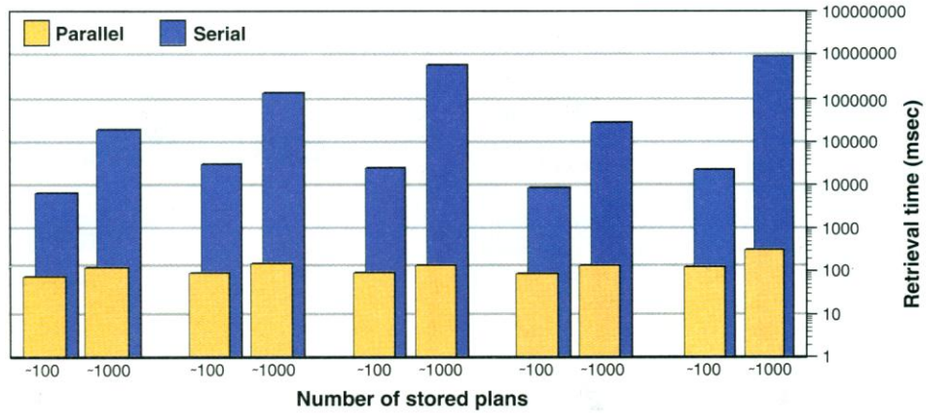
In the past decade, AI research has created important technologies. The annual investment in, and return from, the several thousand existing systems using AI technology is in the hundreds of millions of dollars. Indeed, AI technology has become a part of mainstream software, with applications running from consumer products such as the Tax Cut income tax adviser, the Grammatik grammar adviser (for word processors), and the advanced help system in Microsoft Word 6.0 to expert systems that plan the loading and unloading of cargo ships (Singapore) and the space shuttle refurbishing operations at the Kennedy Space Center. In fact, microprocessor-based AI systems are now running in television sets, cameras, and other low-cost consumer goods.

One feature these successful programs have in common is that they work in well-defined domains in which the systems' information, or knowledge base, is not extremely large. Typically, AI systems produce their answers based on no more than several hundred facts concerning the area of their expertise. Although this is enough for many interesting problems, algorithmic difficulties have prevented the scaling of AI technology to much larger problems that require rapid access to many thousands or even millions of facts. Such very large knowledge bases (VLKBs) are necessary to many applications however, particularly those motivated by the exponential growth of the information resources and needs of our society.

The technology to access information is rapidly becoming acknowledged as a key for addressing many of the nation's most pressing problems. We live in an age in which the availability of information is staggering. Although corporations can maintain their economic and employee records in very large structured databases, traditional database techniques do not work well for many other kinds of knowledge. Scientific endeavors such as the human genome project or NASA's EOSDIS are currently generating massive amounts of information that cannot be easily characterized in the same way as corporate data. Medical records, critical for both personal healthcare and epidemiology research, are swamping hospital computers. And, most ambitious of all, U.S. military intelligence organizations, in a concept almost as intriguing as it is scary, are talking about creating a database nearly ten orders of magnitude greater than any currently existing. This database would contain multimedia information about current events, political data, and the many other types of information of concern to their world view. In fact, the data are so rich in many repositories that "mining" for information relevant to a problem at hand has become a difficult and time-consuming process. To indulge in a currently popular metaphor, the information superhighway is already developing traffic jams. Without some sort of "intelligent" guidance, finding our way through this rapidly expanding information space will become impossible.

Dealing with extremely large amounts of information has long been a challenge to some researchers in the field of AI. For many people, the very phrase "artificial intelligence" conjures up a vision of an intelligent computer that can provide immediate access to vast amounts of information. Such systems, like the HAL 9000 computer from Arthur Clarke's *2001: A Space Odyssey* or the superhuman android, Lieutenant Commander Data, of the television program "Star Trek: The Next Generation" remain squarely in the realm of science fiction, but they are never far from the hearts of many AI researchers. In fact, many early researchers in the field set out to create such programs. Their failures led to the realization that to provide intelligent help in dealing with large amounts of information, an AI system must itself have access to large amounts of knowledge. The AI scientists call this the "knowledge is power" hypothesis, or more simply, "the knowledge principle".

This need for large amounts of very broad knowledge in many applications was the primary motivation leading to the creation of the well-known CYC project (*1*), an effort to build a very large knowledge-base of common-sense information. Over a 9-year span, many researchers at Microelectronics and Computer Technology Corporation (MCC) in Austin, Texas, have developed a conceptual structure for common-sense knowledge (called an ontology) and have populated it with more than half a million facts involving tens of thousands of objects and entities. Although the jury is still out on whether this particular VLKB will succeed, it is clear that VLKBs of at least this order of magnitude are critical to AI's ability to scale up and to have an even broader impact as a technology necessary in solving the nation's most



High-performance AI inferencing. Comparison of the growth (logorithmic scale) of serial and parallel pattern matching as memory size grows. The graph shows five separate knowledge-based queries, used as part of a case-based planning system, and their retrieval times on two separate case bases. Blue bars indicate the retrieval time for a serial system running on a Macintosh IIcx computer with 20 megabytes of memory. Yellow bars indicate the run times of the same queries on a 16384 processor CM2. As can be seen, the parallel retrieval algorithm scales well when compared with the serial method.

The author is in the Department of Computer Science, University of Maryland, College Park, MD 20742, USA.

pressing industrial and scientific problems.

Recently, the use of high-performance computing techniques has led to the development of methods for building and accessing VLKBs with the faster processing and large memories enabled by modern supercomputers. Whereas entering the knowledge base for CYC has taken tens of person-years, these new techniques permit the automatic generation of VLKBs in much shorter times. In addition, new accessing techniques provide searches that are several orders of magnitude faster than serial algorithms for matching complex patterns on relatively unstructured data. Although these techniques may not obviate the need for CYC-like projects (after all, common sense is hard to find, even among people), they open many intriguing possibilities. Some examples are mentioned here:

1) *Large case-based systems*. Instead of solving problems from scratch, systems can solve new problems by analogy to previous solutions (2). Such "case-based" reasoning requires very large memories of previous problem solutions. Building these memories by hand can be an enormously difficult knowledge engineering task. Recent work has shown that large case bases can be automatically generated with AI techniques (3) and accessed extremely efficiently by parallel inferencing techniques (4).

2) *Hybrid knowledge and databases*. Many large corporate and scientific databases can be used to create AI knowledge bases. First, specific information about the domain of discourse is used to encode knowledge about the underlying characteristics and functions of objects in that domain. Following this, traditional database queries are used to create a knowledge base relating specific instances from the database to the more generic AI knowledge. The resulting hybrid knowledge and database can be used to combine searching and inferencing, with supercomputing techniques again providing efficient pattern-matching capabilities that are difficult to encode and inefficient to run in the unaugmented database. This technique is particularly important in applications where old data must be explored in novel ways to see if recently discovered patterns were previously existent in the database (examples include epidemiology and pharmaceutical research).

3) *Software agents*. A recent innovation in AI technology is the creation of intelligent agents to help users explore complex unstructured information, such as that in the millions of documents distributed across the Internet. Although not yet a commercial technology, software agents are expected to become an important mechanism in providing access and navigation aids to the large amounts of information stored in so-called cyberspace. Current Internet agents, for example, provide knowledge-based interface tools for making the net more user friendly (5) and use AI techniques to help filter out vast amounts of irrelevant information (6). A recently started project in my laboratory, for example, focuses on the use of parallel inferencing techniques to provide a basis for creating agents that wander through the network, explore the information residing therein, and process it to create a large, knowledge-based memory for use by interface, search, and filtering agents.

It is still early in the development stage of this exciting new technology, and much important research remains to be done. High-performance computing support for massive knowledge bases requires the exploration of several areas, which are currently receiving only minimal funding (despite the vast budget resources being put into the national information infrastructure). To continue scaling AI, we must explore the use of secondary storage in ways that are amenable to the irregular memory usage patterns of AI's knowledge-base technology (very different from the more regular accesses used in most scientific supercomputing). New inference classes and access mechanisms must be developed for efficiently exploring the ever larger knowledge repositories available to users. New techniques must also be found for mining the data stored in scientific and medical databases, and reasoning mechanisms must be developed for extending software agents technology, tailoring it to the creation and use of large knowledge-based memories. As these techniques become more commercially viable, we can expect to see the next great stride in the use of AI technology in industrial, government, and scientific applications.

References

1. D. Lenat and E. Feigenbaum, *Artif. Intell.* **47**, 1 (1991).
2. J. Kolodner, *Case-Based Reasoning* (Morgan Kaufman, San Francisco, CA, 1993).
3. B. Kettler, J. Hendler, W. Andersen, M. Evett, *IEEE Expert*, **9** (no. 1), 8 (1994).
4. M. Evett, J. Hendler, L. Spector, *J. Parallel Distrib. Comput.*, in press.
5. O. Etzioni and D. Weld, *Commun. ACM* **37**, 72 (1994).
6. P. Maes, *ibid.*, p. 30.

Enterprise-Wide Computing

Andrew S. Grimshaw

For over 30 years, science fiction writers have spun yarns featuring worldwide networks of interconnected computers that behave as a single entity. Until recently, such fantasies have been just that. Technological changes are now occurring that may expand computational power just as the invention of desktop calculators and personal computers did. In the near future, computationally demanding applications will no longer be executed primarily on supercomputers and single workstations dependent on local data sources. Instead, enterprise-wide systems, and someday nationwide systems, will be used that consist of workstations, vector supercomputers, and parallel supercomputers connected by local and wide-area networks. Users will be presented the illusion of a single, very powerful computer, rather than a collection of disparate machines. The system will schedule application components on processors, manage data transfer, and provide communication and synchronization so as to dramatically improve application performance. Further, boundaries between computers will be invisible, as will the location of data and the failure of processors.

To illustrate the concept of an enterprise-wide system, first consider the workstation or personal computer on your desk. By itself it can execute applications at a rate that is loosely a function of its cost, manipulate local data stored on local disks, and make printouts on local printers. Sharing of resources with other users is minimal and difficult. If your workstation is attached to a department-wide local area network (LAN), not only are the resources of your workstation available to you but so are the network file system and network printers. This allows expensive hardware such as disks and printers to be shared and allows data to be shared among users on the LAN. With department-wide systems, processor resources can be shared in a primitive fashion by remote log-in to other machines. To realize an enterprise-wide system, many department-wide systems within a larger organization, such as a university, company, or national lab, are connected, as are more powerful resources such as vector supercomputers and parallel machines. However, connection alone does not make an enterprise-wide system. If it did, then we would

The author is in the Department of Computer Science, University of Virginia, Charlottesville, VA 22903, USA.