MULTI-COMPUTER ARCHITECTURES FOR ARTIFICIAL INTELLIGENCE

Toward Fast, Robust, Parallel Systems

LEONARD UHR

MULTI-COMPUTER ARCHITECTURES FOR ARTIFICIAL INTELLIGENCE

Toward Fast, Robust, Parallel Systems

LEONARD UHR

Computer Sciences Department
University of Wisconsin, Madison

A Wiley-Interscience Publication

John Wiley & Sons

New York Chichester Brisbane Toronto Singapore

The author greatfully acknowledges the use of the following figures: Figure 1.2 courtesy of C. E. Shannon and J. McCarthy, eds., *Automata Studies*, Annals of Mathematics Studies No. 34, Copyright © 1956, 1984 renewed by Princeton University Press; Figure 3, p. 134, reprinted with permission of Princeton University Press; Figure 6.2 courtesy of Hwang and Briggs, Computer Architecture and Parallel Processing, 1984, McGraw-Hill, (reproduced with permission).

Library of Congress Cataloging in Publication Data:

Uhr, Leonard Merrick
Multi-computer architectures for artificial intelligence.

"A Wiley-Interscience publication."
Bibliography: p.
Includes index.
1. Artificial intelligence—Data processing.
2. Parallel processing (Electronic computers)
3. Computer networks. 4. Computer architecture.
I. Title.
Q336.U37 1987 004'.35 86-15746
ISBN 0-471-84979-0

MULTI-COMPUTER ARCHITECTURES FOR ARTIFICIAL INTELLIGENCE

PREFACE

MULTI-COMPUTER ARCHITECTURES FOR ARTIFICIAL INTELLIGENCE

This book is directed toward computer scientists, along with other research scientists and engineers, who are interested in artificial intelligence on the one hand, and/or parallel computer architecture on the other hand. It is also designed to inform the much larger group of people intrigued by the great explosion of possibilities today's phenomenally rapid progress in microelectronic "very large scale integration" (VLSI) opens up, and in the profound potential promise of artificial intelligence. It concentrates on the overall topology of a multi-computer (rather than the design of the individual computer), and on the development of powerful really intelligent systems (rather than programs for specific small problems).

This book has three major purposes:

1. It attempts to give, with a detailed survey and a large number of examples, a clear, concise picture of the great variety of multi-computer architectures already built, already designed, now being designed, and feasible—in the immediate and in the more distant future.
2. It describes the range of problems that confront AI, and sketches out some of the major approaches being taken by AI researchers. Here it focuses on and emphasizes the crucial importance of robust, flexible, extremely fast (hence highly parallel) systems.

Such systems are necessary to handle the enormous variability and stringent real-time demands of the real-world problems of perceiving often greatly distorted and poorly sensed objects and concepts, remembering and making inferences in ill-formed domains, and controlling and coordinating an intelligent robot that can do all these things. These entail a number of seemingly simple, mundane problems: for example, to recognize and describe people and objects in a photo; to talk and think about all sorts of "ordinary" things; to navigate the streets and find and get (buy, beg, borrow) things that satisfy one's needs. These are problems that good, general AI systems must handle to achieve real power and real intelligence (as opposed to the simplified and rigid toylike problems for which AI systems are too often developed).

3. It also presents judgments and opinions as to the most promising multi-computer architectures for AI. It suggests the most promising architectures in terms of (a) meeting AI's ultimate long-term goals of general, fast, robust systems; and (b) the shorter-term development of special-purpose systems that have important practical applications. These architectures include a number of exciting multi-computer structures of which the AI community is largely unaware, and also several suggestions of new and promising structures that appear to be worthy of thorough development and evaluation.

This book examines the present state of parallel multi-computers and of artificial intelligence, but it emphasizes the future. Most AI programs have been designed to handle a specific problem, using serial algorithms for the serial computer. These traditional serial programs are surveyed; but a variety of potentially far faster and more powerful parallel approaches are suggested and explored.

It is almost certainly the case that many of the most interesting and promising parallel architectures (both hardware and software) have not yet been built, or even designed. This book—by examining the variety of feasible multi-computers alongside possible parallel formulations of AI problems—attempts to stimulate an integrated approach to the development of new software/hardware systems, where each is appropriate to the other. The structure of the processes to be executed can, once understood, suggest hardware structures that will execute them with speed and efficiency.

Today's micro-electronic technologies already make possible, and feasible, highly parallel networks with many thousands, or even millions, of computers all working together on the same task. Systems of this

size will open up an intriguing range of new possibilities, for physics, chemistry, meteorology, engineering, and computer science. Without computers that are enormously large and highly parallel, true (artificial or natural) intelligence is impossible.

Today we are witnessing the beginnings of a major surge of interest in parallel computing. Very large parallel multi-computer networks appear to many people to offer the only possible way to make today's conventional serial single-CPU (central processing unit) Von Neumann computers substantially more powerful. This is because conventional computers are rapidly reaching the point where they can no longer be made faster; they are approaching the "speed of light barrier." The basic component of a computer—the switch, today realized with transistors—can now be made so incredibly small, hence fast, that the time taken for signals to move along the wire joining two transistors is on the verge of becoming the dominating factor in determining the speed at which computations can be effected.

Artificial intelligence requires very fast and very powerful multi-computer hardware structures for real-world problems where perception, decision making, remembering, inferring, and the other components of intelligence must work within extremely stringent real-time constraints. This will be the case whenever the AI system (e.g., an intelligent robot) must recognize, move to, gain control over, and/or interact with, complex real-world objects as they move about and change.

Therefore AI researchers are increasingly turning to parallel processes. A few have suggested a handful of parallel architectures; but almost nothing has been written that examines, from the point of view of their appropriateness for AI, the many different kinds of multi-computer networks that have been and are increasingly being explored and developed by the larger group of computer architects. This book attempts to fill that gap.

OVERVIEW

The Introduction poses the problem of very fast real-time perception.

Chapter 1 examines the underlying model of a general purpose computer that it embodies.

Chapter 2 describes the traditional single-CPU serial computer, and some of the extensions (many of which are additions of parallel hardware within the single CPU) that have been made to increase its speed and its power.

Chapter 3 presents the great advances in micro-electronics that are beginning to make feasible very much larger, far more parallel, computers.

Chapter 4 introduces the enormous variety of possible topological structures for parallel multi-computers (essentially, all possible graphs), with attempts to put several of these possibilities into a common framework.

Chapter 5 explores the major problems (with which researchers are today just beginning to grapple) in developing fast, efficient, appropriately structured operating systems for highly parallel, closely coupled multi-computers.

Chapters 6-13 examine the great variety of parallel architectures that are running, built, designed, proposed, or possible, including:

a large number of different topologies that are being used or investigated, for both (pseudo-) complete graph connectivity (bus, ring, crossbar, reconfiguring network) and point-to-point connectivity (e.g., line, array, N-cube, pyramid, tree, augmented tree, compounds of clusters, data-flow, heterogeneous structures), along with examples of multi-computers that are based on each;

possibilities for developing more general purpose yet appropriately structured architectures;

two major (overlapping) classes of multi-computer: (a) networks that attempt to be general purpose over a great variety of (or all possible) programs, and (b) more or less specialized algorithm/process/task-structured architectures.

Chapter 14 briefly describes the human brain, both (a) as an existence proof that massively parallel structures can succeed at extremely fast real-world real time processing, and also (b) as a fruitful source for design ideas.

Chapters 15-21 present the major areas of artificial intelligence, and the range of different approaches that researchers are taking—with emphasis on parallel approaches. These include:

pattern perception, computer vision, image processing;

structuring, representing, and accessing information, whether symbolic, linguistic, iconic, or perceptual;

recognizing, analyzing, and understanding speech and language;

thinking: problem solving, theorem proving, decision making, "expert" systems;

robot motor control and coordination;

learning;

integrative systems that combine intelligent processes.

Chapters 15-21 also present the multi-computer network architectures that appear to be the most appropriate for each AI function.

Chapters 22-24 summarize the different architectures examined, evaluate their strengths and weaknesses, and suggest several architectures that appear to be the most promising for a total intelligent system.

The variety of architectures examined in this book is greater than can be found in any other publication of which I am aware. I have tried to choose the most representative, the most interesting, and the most powerful; but these are best considered as examples. The possibilities for the future, both of parallel computers and of artificial intelligence, are so great that today's best understanding will surely be improved upon substantially.

LEONARD UHR

Madison, Wisconsin
February 1987

CONTENTS

PART ONE. BACKGROUND INFORMATION AND BASIC STRUCTURES

INTRODUCTION: KEY ARTIFICIAL INTELLIGENCE PROBLEMS, AND ARCHITECTURES 1

1. UNIVERSAL TURING MACHINES AND THE THEORETICAL BASIS OF COMPUTERS 11
2. CONVENTIONAL GENERAL-PURPOSE SINGLE-CPU SERIAL COMPUTERS AND SUPER-COMPUTERS 20
3. THE PRESENT, IMPENDING, AND FUTURE TECHNOLOGICAL POSSIBILITIES 26
4. THE BASIC GRAPH TOPOLOGIES FROM WHICH MULTI-COMPUTERS CAN BE CONSTRUCTED 36
5. THE BASIC HARDWARE AND SOFTWARE COMPONENTS OF MULTI-COMPUTERS 43

PART TWO. PARALLEL AND PROCESS-STRUCTURED MULTI-COMPUTER ARCHITECTURES

6. LINKING SMALL NUMBERS OF COMPUTERS INTO PSEUDO-COMPLETE GRAPHS 56
7. PIPELINES AND ASSEMBLY LINES 77
8. VERY LARGE NETWORKS STRUCTURED INTO ARRAYS 92
9. AUGMENTED ARRAYS; PYRAMIDS; TOWARD AUGMENTED PYRAMIDS 110
10. POINT-TO-POINT LINKED TOPOLOGIES FOR LARGE MIMD MULTI-COMPUTERS 124

11. GOOD (AND A FEW OPTIMAL) TOPOLOGIES FOR MULTI-COMPUTERS 134
12. TOWARD VERY LARGE MIMD NETWORKS WITH POTENTIALLY MILLIONS OF COMPUTERS 148
13. TOWARD MORE GENERALLY USABLE AND/OR MORE ALGORITHM-STRUCTURED ARCHITECTURES 162

PART THREE. ARTIFICIAL INTELLIGENCE: MAJOR APPROACHES; INDICATED ARCHITECTURES

14. THE HUMAN MIND/BRAIN, AND POSSIBILITIES OF PARALLELISM IN EACH AI FUNCTION 181
15. PERCEPTION: IMAGE PROCESSING, PATTERN RECOGNITION, COMPUTER VISION 193
16. STRUCTURING AND ACCESSING SYMBOLIC, LINGUISTIC, ICONIC AND PERCEPTUAL INFORMATION 214
17. SPEECH AND LANGUAGE: ANALYSIS, RECOGNITION, UNDERSTANDING 225
18. PROBLEM-SOLVING, THEOREM-PROVING, DECIDING; "EXPERT SYSTEMS" 232
19. ROBOT MOTOR CONTROL AND COORDINATION 247
20. LEARNING 254
21. WHOLISTIC SYSTEMS THAT INTEGRATE SEVERAL INTELLIGENT PROCESSES 270

PART FOUR. THE PRESENT STATE; MAJOR PROBLEMS; MOST PROMISING ARCHITECTURES

22. A CRITICAL LOOK AT ARCHITECTURES PROPOSED FOR AI 281
23. SUGGESTIONS AS TO THE MOST PROMISING ARCHITECTURES 287
24. PARALLEL ARCHITECTURES SUMMARIZED, AND SEVERAL KEY ISSUES 293

REFERENCES 310

INDEX 341

KEY ARTIFICIAL INTELLIGENCE PROBLEMS, AND ARCHITECTURES

This book explores three closely related problems, examining the present situation and the range of possibilities for the immediate and especially for the more distant future: A. designing and building highly parallel multi-computers (networks of computers that all work together on the same problem) that can with speed and efficiency execute substantially larger programs than can today's computers; B. developing hardware/software systems of computers and programs that are truly intelligent; C. designing these computers and programs jointly, so that each is appropriate to the other.

The continuing steady and rapid increase in the potential size of multi-computers that can be built makes feasible a great variety of possible architectures—along with a formidable set of difficult problems that must be solved to use them effectively. For, potentially, a multi-computer can be built using any possible graph topology.

Artificial intelligence is an enormously difficult endeavour. Ultimately it attacks the age-old problem of philosophy and psychology: What is mind/brain, and how does the mind/brain know, understand, and cope with its world? Today we tend to break this problem down into sub-problems: What is perception, cognition, reasoning, concept formation, remembering, motor behavior, learning, discovery, creativity? We tend to choose a particular sub-sub problem to start with—for example, recognizing two-story houses, following stories about restaurants or terrorists, robot assembly, chess—and program in great globs of ad hoc knowledge about these tasks. Each of these functions must be

generalized, and all must be integrated, to address the crucial overall question: What is intelligent thinking?

Very flexible, robust, and powerful programs are needed to handle AI problems in a reasonably general way, coping with the enormous range of unanticipated variations. These programs will need highly parallel multi-computers, simply because there are theoretical limits to the power and speed that can ever be achieved with a single conventional computer, even an ultimate super-super version of what today are called "super-computers." And the human mind/brain—which is the only successful intelligent system discovered to date—achieves its great power, robustness, and speed because it is a highly parallel multi-computer.

It is important that the AI community become aware of the many types of highly parallel networks that have already been built or designed, and of the far greater number that are now possible, and develop AI programs that address the fundamental AI goal—the ability to carry out very powerful, general, robust, fast, intelligent thinking in real time. Only then will people be in a position to explain AI's problems and needs, and to ask for multi-computers that fill these needs.

ARTIFICIAL INTELLIGENCE TASKS AND GOALS, AND NEEDED SPEED

The following briefly summarizes key central long-term AI goals for which both highly parallel algorithms and highly parallel multi-computers appear to be absolutely necessary. The human mind/brain's incredible speed sets these levels of performance. It is important to emphasize that the functioning mind/brain also serves as a crucial existence proof that these goals can be achieved.

A. Programs for visual perception must be able to process continuing streams of complex—often severely degraded and distorted—real-world images (each 256 by 256, 512 by 512, 1024 by 1024, or larger, input every 20 to 50 milliseconds), recognizing and describing, and tracking, the salient and significant objects. They must also be able to recognize up to a hundred or so complex objects within a second or less, as when reading a book at the rate of 300 to 1000 words a minute, or when recognizing a complex scene with many complex objects.

B. Systems that, often in situations of ambiguous and/or incomplete information, give "expert" advice, make decisions, choose, diagnose, reason, infer, or solve problems, should (as people do: that

is, usually but not always) arrive at appropriate responses within one to five seconds, at most.

C. Speech and language recognition systems must handle casual, often heavily accented and/or poorly enunciated, continuous speech that is spoken by a variety of different speakers, at the rate of 2 to 4 words per second.

D. Semantic memory networks should (often but not always) be able to access correct, or appropriate and cogent, answers or responses to relatively unconstrained, often poorly phrased or ambiguous, natural language sentences, in 1 or 2 seconds or less (the speed that is routinely achieved by human beings engaged in a conversation), or at most (but only when we are in "deep thought") in 5 or 10 seconds.

E. Robot perceptual-motor systems must make use of information perceived through several sensory channels (typically, vision, hearing, and touch) in order to coordinate and control the motions of the robot's various appendages while the robot is moving at realistically fast speeds (ideally, 2 to 100 miles per hour or, for certain applications, even faster, as when, by coordinating hundreds of interacting muscles, a human being crawls, walks, bats a ball, drives a car, flies a plane, or tries to stop a moving bullet or missile).

F. A system that integrates several of the above functions (e.g., a robot that interacts with a dynamic environment, using vision, speech, and touch to guide its behavior, and also deduces appropriate responses) should be capable of appropriate continuing interactions with moving objects and people in its environment in real time (that is, at the speeds with which these real objects and people act).

G. Any of these systems should be able to learn and adapt, as a function of its experiences; the basic experiences from which the system learns must be of the kinds described earlier; the system should learn each specific bit of information in seconds or less; and the learning should achieve good levels of performance fast enough for the system to survive and function effectively.

The human brain achieves these enormous speeds even though its basic "instruction cycle time" is roughly 1-2 milliseconds (the time needed for one neuron to fire another neuron across the synaptic gap between them). That is, the brain's neurons are millisecond devices, in stark contrast to today's computers, whose basic instruction cycle times

are on the order of microseconds, or even a few nanoseconds. Thus a computer's electronic circuits are four to eight orders of magnitude faster than the brain's neuronal circuits; yet computers are, when attempting (but not yet succeeding) to handle intelligent processes, roughly four to eight orders of magnitude slower than are brains.

In the 1-2 milliseconds taken to cross a single synapse, the brain accomplishes tremendously impressive feats in both speed and performance—but only because enormous numbers of the brain's neurons are all working in parallel. Even with their much faster basic cycle time, computers cannot possibly work as well or as fast without massive parallelism. For example, the mind/brain recognizes complex objects in a few hundred milliseconds; this means that the serial depth of object recognition is at most a few hundred. But today's first attempts at computer vision programs, when executed on a serial computer, need serial sequences of billions of instructions, and rough estimates suggest that trillions may be needed for successful programs.

THE NECESSITY OF HIGHLY PARALLEL MULTI-COMPUTERS FOR TRUE, ROBUST INTELLIGENCE

The Great Complexity of Real Intelligence, and of Programs Capable of Real Intelligence

This book focuses on artificial intelligence's basic long-term goals: to develop software/hardware systems (they might best be called something like "information processing structures," rather than "programmed computers") that are really intelligent in at least the variety of ways that human beings are intelligent. True intelligence, whether natural or artificial, entails real understanding of the world—of the organisms and objects that inhabit it, of the symbols and ideas that attempt to cope with it and explain it—and of oneself and how to interact purposefully with that world.

This understanding is never complete (indeed it seems likely that there is no such thing as complete 100% understanding); but it is relevant and sufficient, and capable of changing and growing. The perceiver really understands what houses, chairs, and people are. This is a deep understanding that makes possible powerful recognition over extreme distortions and relevant reactions, whether verbal descriptions or complex actions. The language understander really comprehends the meaning of the comment, directive, story, or book. The chess master, master chef, auto mechanic, race car driver, short order cook, mathematician,

computer architect, engineer, bartender, brew master, and brewery designer all really understand their own domain, from their own perspective. The "understandings" of a Ferrari by the sales-person, driver, mechanic, mechanical engineer, designer, physicist, psychologist, sociologist, economist, moral philosopher and metaphysician will differ greatly. A truly intelligent AI system need not have all of these levels and perspectives of understanding of all possible things (indeed no single human being does, and that is probably impossible), but it must be capable of achieving (by learning) any of these.

The Great Size of the Hardware Multi-Computer Needed for Intelligence

Parallel algorithms, and very large multi-computer architectures that handle these algorithms efficiently, are absolutely necessary for real time execution of these kinds of large AI systems—to make them work on real-world, as opposed to toy, problems. It seems likely that to perform with the speed and power of the human brain (with its many billions of neurons, each of as yet undetermined complexity) networks of many thousands, and probably millions and possibly billions, of computers will be needed, although each individual computer may be relatively small.

At one extreme, we can contemplate enormously large numbers of very simple processors (e.g., simple boolean logic gates, or threshold logic processors, or idealized neuron-like devices, or simple and basic 1-bit computers). At the other extreme, much smaller numbers of extremely powerful computers may be preferable, especially where massively parallel processes and extremely fast speeds (which are vital for visual perception and robot control) are not so crucial. And a large number of possibilities exist in between.

PARALLEL ARCHITECTURES BEING CONSIDERED BY THE AI COMMUNITY, AND THEIR PROBLEMS

AI researchers interested in problem solving have focused their attention on three major types of hardware multi-computers:

A.1. Specialized computers and small networks to execute expert production systems or/and to do logical inference;

A.2. Larger networks to handle the general AI problem, possibly by assigning different functions to different groups of computers;

A.3. Very large networks to handle semantic memory searches.

The closely related image processing, pattern recognition, and computer vision communities have concentrated on:

B.1. Small networks to do an essentially divide-and-conquer examination of the large image array;
B.2. Very large synchronized arrays of processors to execute in parallel iterated local operations over the entire array of image pixels (picture elements).

It appears, unfortunately, that attempting to parallelize today's "expert" production systems and logical inference systems can give at best only very small amounts of speedup—on the order of 2 or 3, or possibly 10 or 20 at most. These programs, as they are coded today, are almost always designed to execute serial algorithms in a serial manner.

It is exceedingly difficult to develop parallel algorithms. Very possibly this is because our "conscious" thinking is mostly serial; although our not-conscious thinking is mostly parallel—but inaccessible to conscious thought. Another major obstacle to increasing the speed of today's programs is that it is far more difficult, and often impossible, to parallelize a serial algorithm after it has been coded in a serial fashion than to attempt to devise a parallel algorithm in the first place. Parallel formalizations appear to be possible for both production systems and logical inference systems, and there is reason to hope that they can offer greater speedups, possibly as great as two to four orders of magnitude (see Chapter 18).

A large number of networks of completely independent computers have been built with 8 or 16 processors. Several systems have been designed, and a few built, with from 50 to 256. Most of these (described in Chapter 6) link computers via a pseudo-complete graph. (A complete graph has each node linked directly to every other node.) This reflects many people's feelings that complete connectivity is the ideal. A number of structures have been proposed for multi-computers with several thousand, or even more, processors (see Chapters 7-13). Some of these use a tree (a graph with no cycles, e.g., Browning, 1980; Mago, 1980), a tree with some extra links (e.g., Shaw, 1982), or a binary N-dimensional cube—that is, an N-dimensional array with nodes (computers, processors) only at its corners (e.g., Seitz, 1985). However, the topology of such networks too often has no particular relation to the structure of the programs they will execute. And the speedups of