PARALLEL PROCESSING COMPUTER-ARCHITECTURE FOR INFORMATION RETRIEVAL (ARCHITECTURE D'UN ORDINATEUR A CALCUL PARALLELE ORIENTEE VERS LA RECHERCHE DE L'INFORMATION)

T. Radhakrishnan
Computer Science, Concordia University, Sir George Williams Campus
Montreal, H3G iM8

V. Rajaraman
Computer Science, I.I.T. Kanpur
India

ABSTRACT

Architectures involving a single central processor, controlling a number of parallel processors have been proposed. In this paper, we discuss a hierarchical computing system with a number of specialized parallel processors to perform the information retrieval operation. Such an architecture is well suited for changing demands and reconfiguration is simpler. It is suitable for very large data bases and will find its role in national information networks. (Des architectures fournissant un processeur central unique qui controle plusiers processeurs parallèles ont été proposées. Dans ce papier nous discutons un système hierarchique de calcul qui comprend plusiers processeurs parallèles specialisées dans les opérations de recherche d'information. Une telle architecture est particulièrement adaptée à des modifications et la réconfiguration est plus simple. Elle est particulièrement utile pour des grandes bases de données et trouvera son rôle dans "des réseaux nationaux d'information".)

COMPUTER ARCHITECTURE FOR I.R.

1. INTRODUCTION

The advent of integrated circuits and the decreasing cost of digital hardware have motivated the design of computer systems with multiple processors. Such architectures are considered to suit certain types of problems and hence are "well matched" for that application (Barnes 1968). In this paper, we consider an architecture that is well matched for information retrieval. Information retrieval from a data base is a process rich with parallelism. Suppose there is a query Q, and n files F1,F2,...Fn are to be searched to find the response to Q. The search in these n files can be done in parallel and quite often these parallel searches are mutually independent. In fact, this partitioning of files can be extended to an elementary level. Then, what we need is a true associative memory of sufficient size, in which each memory cell is a "processor" in some sense. However, the cost of such large size associative memories and the lack of flexibility make them impractical with the current technology.

Information retrieval is an iterative process (Figure 1). The user generates a query and it is expressed in a suitable "query language". This input is processed by the "query processor" which uses some dictionaries. Synonym dictionary and classification or hierarchical dictionaries are some examples (Salton 1968). It is possible to keep the "search files" in an encoded form, in which case the search-terms are also to be encoded using the same coding method (Files 1969). Thus the output from the query processor is a set of search-terms which is in a suitable form for the search process. In determining these search-terms there could be feedback from the user as indicated in Figure 1. A "search file" is a collection of indexed and possibly coded documents. As the retrieval using the entire document text in its natural form is expensive, such indexing and coding are unavoidable. Collection of these indexed documents called "search data base" is partioned into a number of files. This partitioning could be based on the logical characteristics of the data base, usage patterns, physical storage characteristics, or combinations of them.

The search terms form an imput to a retrieval function F(..) along with the search files. The output from the retrieval function is a set of pointers, possibly null, to the relevant documents, whose texts or descriptions could be found in the "document-file". In practice, the complexity of such retrieval functions vary considerably. It could be a simple pattern match, a threshold function, or a complex structure-match (Heaps 1971).

The document-file contains the texts or references to the document-texts. The output of the retrieval function is used to selectively display the "relevant" documents to the query. The

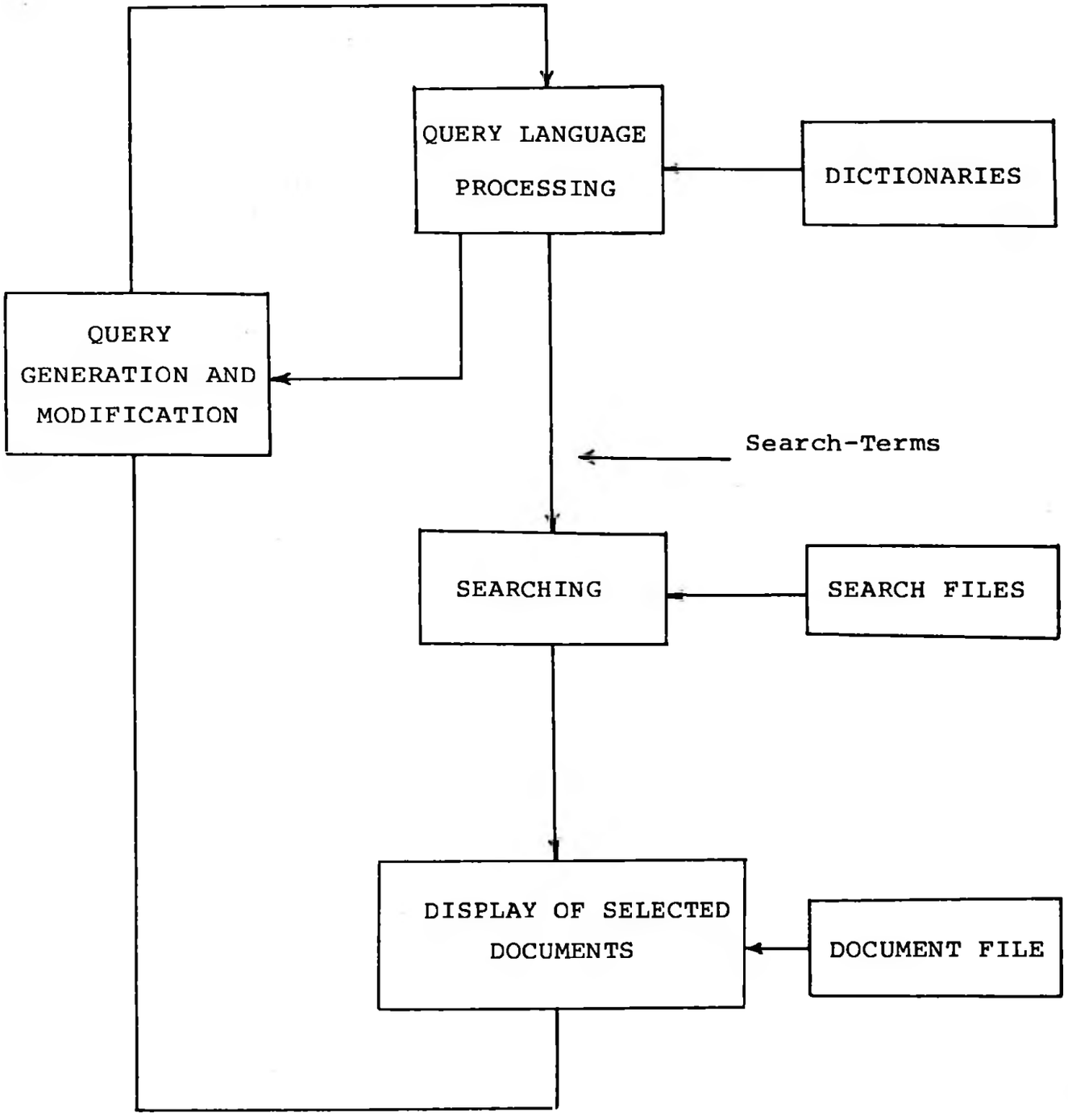

Fig.1

Information Retrieval Process

display mode depends on the hardware available with a system. It could be a hard copy output on printer, microfilm or softcopy display on CRT screens. There is a possible user feedback at this stage as shown in Figure 1. If the selected documents are not what the user expected, he might reformulate his query and start the next iteration.

2. RETRIEVAL PROCESS:

The retrieval process involves three distinct stages, namely, query processing, file searching and the display of selected documents. If the feedback path in Figure 1 is cut (which implies that every iteration of a query is considered as a new task), a top-down representation is possible as shown in Figure 2. We notice that stage-1 and possibly stage-3 involve interaction with users in an interactive system. Hence, they are good candidates for time-sharing. Bulk of the computations in the retrieval process, occurs in stage-2 and quick response is possible if the time spent in this stage can be minimized. This depends very much on the retrieval function, its implementation, the number of processors available for parallel processing, and the file partioning.

3. COMPUTER ARCHITECTURE

The proposed computer architecture has a central "main computer". It is timeshared by many users and all the communications (input/output) with user terminals are done under the control of this machine. In other words, the stage-1 and stage-3 functions of the hierarchical process shown in Figure 2 are done by the main computer. The stage-2 function is realized by a set of parallel processors P1,P2, P3...Pn (see Figure 3). Each Pi is associated with some local memory and a processing unit. On one side, these Pi's are connected to the main computer through the communication buses and on the other side to the search files through the input buses. Every parallel processor is capable of accessing any of the search files with the help of the selector switch. A partitioned search-file SFi is connected to an input (to the Pi's) bus and the information stored in this file is continuously available on the bus. It could be read simultaneously by more than one Pi and the Pi's will be programmed to know when to start and stop reading from the bus. Each Pi has a program stored in its memory to implement the retrieval function. This could possibly be different for different Pi's. Loading such programs into the memory of the selected Pi will be done by the main computer under appropriate mode selection.

The two-way communication between the main computer and Pi's requires exchange of "control commands" and "data". Suitable implementation of these protocols permits asynchronous parallel

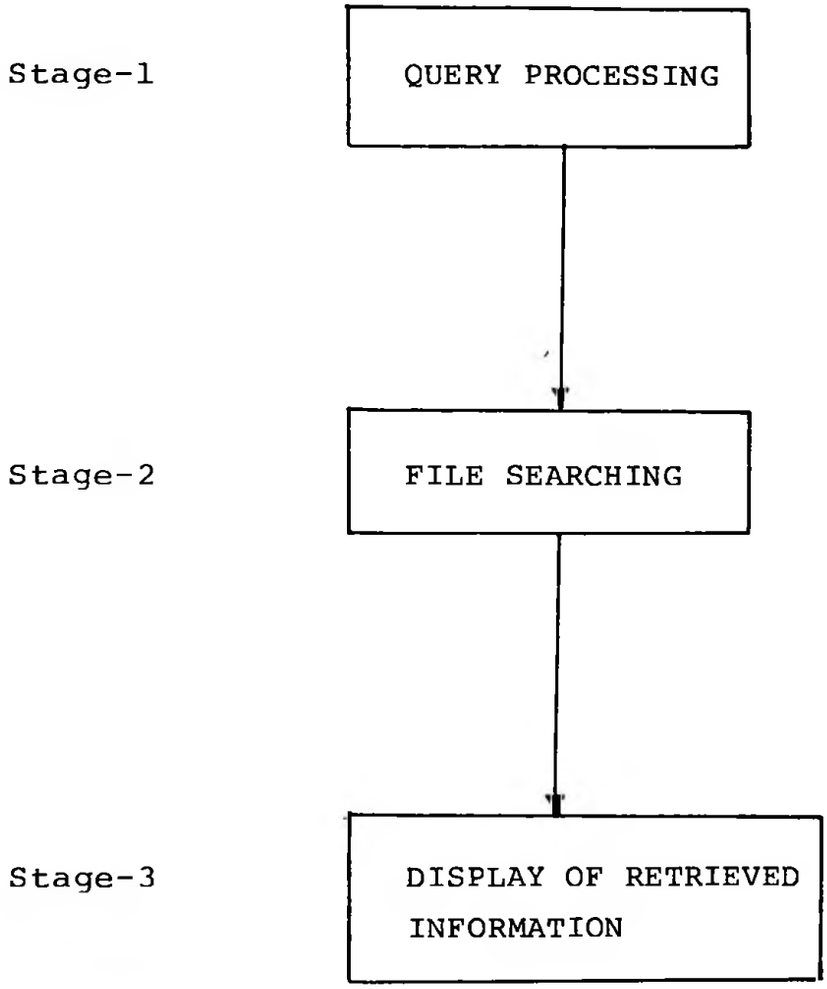

Fig. 2

Top-down representation of Retrieval Process.

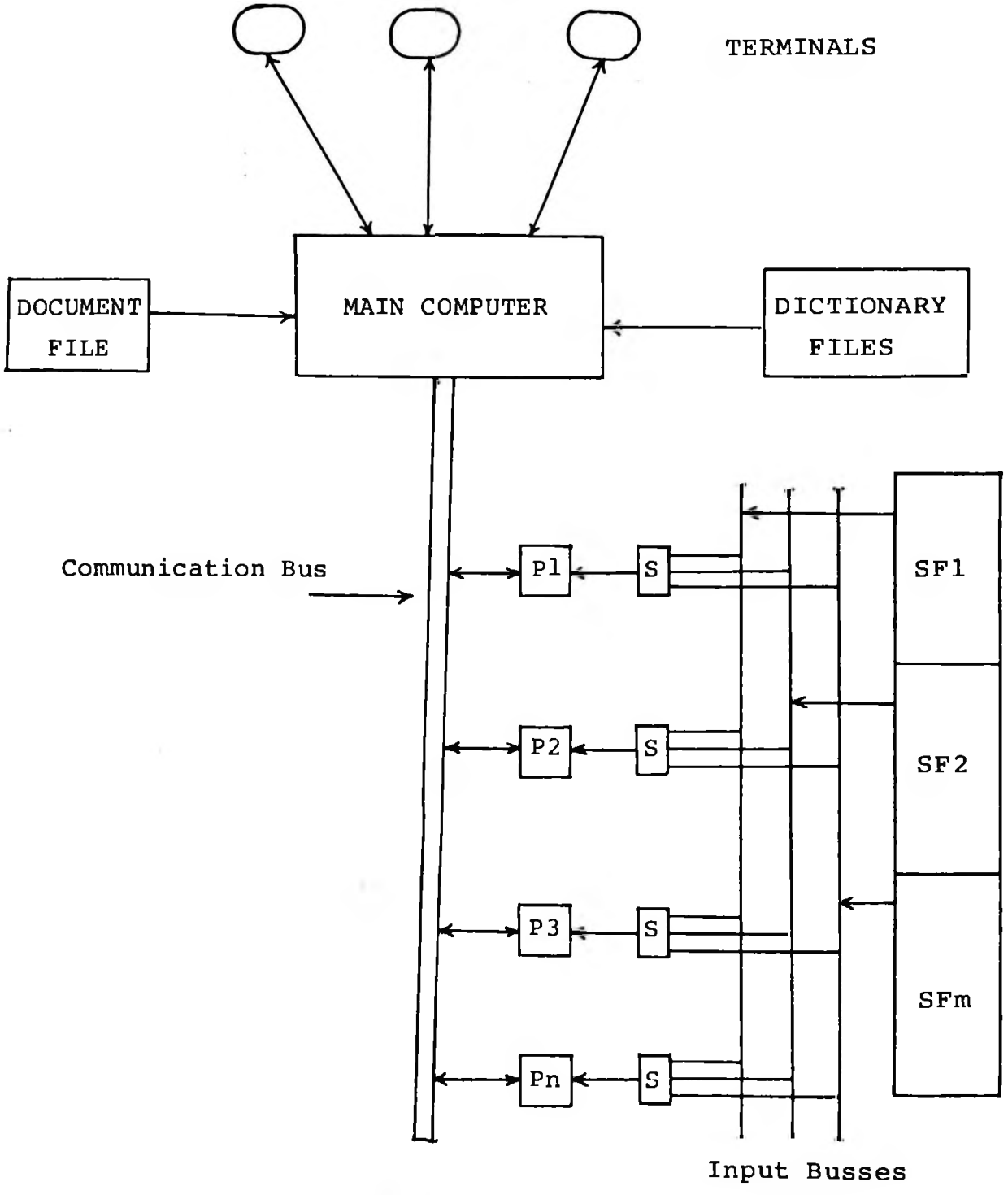

Fig. 3

Computer Architecture for Information Retrieval.

P1, P2, P3, ... Pn	Processors
S	Bus Selector or Switch
SF1, SF2, ... SFm	Partitioned Search Files

operation among the parallel processors. This implies that different Pi's may have different speeds of operation and even of different technology. New processors can be added when the computational need increases or existing Pi's can be detached when the demand is too low. A changing demand can be served using optimum resources without any degradation in the performance level. The choice of instruction set, memory capacity and CPU design of a typical Pi are discussed in (Radhakrishnan 1971). It is interesting to note that with the development in LSI technology more complex Pi's can be added, to operate in parallel with existing Pi's.

4. SOME APPLICATIONS

Two types of information retrieval systems can be conceived: (i) Batch processing of queries (ii) on-line interactive retrieval. Both these types have definite advantages and each is suitable in its own context. One important criterion for on-line retrieval systems is the need for quick response, in the order of few seconds. Any realistic data base is sufficiently large and providing such fast response needs high computational power. Further, centralized information storage and retrieval like CAN/OLE (Canadian On-Line Enquiry at Ottawa) requires communication facility. Set of such information banks can be interconnected to provide nation-wide information networks. Centralized databases and national information networks have many advantages (Overhage 1965). In such cases, the demand for computational power in the retrieval process is stringent and depends on the following factors.

- Number of terminals serviced and the rate of queries generated
- Maximum permissible delay in getting the retrieval output
- Volume or size of the data base
- Characteristics of the queries (broad or specific).

Such information networks have to be failure tolerant and highly reliable. In the proposed architecture failure of a parallel processor can at most degrade the performance and reconfiguration would be straightforward.

5. SIMULATION

The proposed architecture had been simulated at the system level. The inverse relation between the number of processors and the response time for a given fixed condition was studied. Any increase in response time is bought at the cost of adding new processors which also means the increase in cost. By assigning costs to the processors an optimal system configuration could be achieved (Radhakrishnan 1971).

An important aspect of simulation is the characterization of the various conditions. This involves parameters to characterize the queries (mean and probability distribution), the processor (mean processing time and distribution), and to characterize the physical storage devices (mean access time and distribution). It is also possible to study the waiting time of processors or the busy condition of buses. Such simulation could be a useful tool for the system design when the operating conditions are suitably characterized.

6. CONCLUSION

A computer architecture, well matched to the information retrieval applications, is proposed. The basic approach had been to isolate the I/O bound and compute-bound portions of the retrieval operation. The architecture is modeled to match each of them. This organization, being modular, is well suited for changing load conditions and more reliable. Attempts are made to build a content addressed memory using disc systems (Minsky 1972). Unlike most of them, the architecture proposed here is independent of the technology and physical characteristics of the storage devices. The low cost of CPU's and memory (in limited size) with integrated circuits should make this architecture a viable design.

REFERENCES

Barnes, G.H. et. al. 1968. The Illiac IV Computer. I.E.E.E. Transactions on Computers (17) p. 746.

Files, J.R. and H.D. Huskey. 1969 An information retrieval system based on superimposed coding. Proc. of Fall Joint Comp. Conference. V. 35: p. 423.

Heaps, H.S. 1971. Criteria for Optimum Effectiveness of Information Retrieval Systems. Information and Control 18(2): 156-166.

Minsky, N. 1972. Rotating storage devices as partially associative memories. Proc. of Fall Joint Comp. Conference V(41): 587-596.

Overhage, F.J. and R.C. Harman 1965. INTREX Planning Conference 1965. M.I.T. Press.

Radhakrishnan, T. 1971 SYstem design of a parallel processing computer for information retrieval. Ph.D. Thesis submitted to the Department of E.E. I.I.T. Kamput, India.

Salton, G. 1968 Automatic Information Organization and Retrieval McGraw-Hill.