PANEL ON

PARALLEL INFERENCE MACHINES

W. Bibel

Institut fur Informatik, TU Munchecn, Postfach 202420, D-8000 Muncfaen 2

The quest for ever more powerful computers has bumped up hard against the limits imposed by nature such as speed of light and electrons. However, scientists and industry agree that there is still a great potential for further speed-up by distributing computations among many processors rather than a single one. This is apparent for problems that can easily be broken down into many independent parts such as those to be tackled in graphics, signal processing (which includes radar, speech and vision analysis), structural analysis, fluid-flow dynamics, particle physics, and many others. First experiences with the new breed of parallel computers justify this optimism.

It is less obvious whether a multi-processor would bring a significant improvement in performance for problems such as inferencing. Finding a correct chain of inferences requires searching through a space of different possible chains, a problem known to be hard (NP-complete) and requiring exponential recources in worst cases according to our present knowledge. Therefore one might argue that one thousand processors would provide relatively little improvement over a single one in worst (exponential) cases. From a more practical point of view, experiments seem to indicate that the possibilities for exploiting parallelism in rule-based systems might be rather limited.

There are more problems of detail arising in an attempt to parallelize inferencing. The different possible chains in the search space partially coincide in a way not known at compile time hence making it difficult to break the whole task down into many parts and distribute these in a well-balanced way. Also (but not only) because of this overlap it seems attractive to exploit the parallelism inherent in each of these chains. But at present it is not clear at all whether this really pays since it might cause too much overhead in communication and anyway might have only a marginal effect in view of the more severe problem of NP-completeness mentioned above.

Some even put forward reasonable arguments to the effect that inferencing in practice is not needed at all. Part of it could be substituted by exploiting data-base techniques even in the presence of recursion. Or, machine reasoning could be founded on episodes from the past stored in a massively parallel memory rather than on rules and facts thus leading to a memory-based reasoning with a parallelism radically different from the one discussed above; with this remark we scratch on the current discussion about connectionism.

Much of what was pointed out above reflects the spirit of classical reasoning or even more specifically Horn-clause (or production-rule-based) reasoning. As we know human reasoning has many flavors that might still cause (at least practical) problems with their integration into this classical deduction scheme. Non-mono tonic, probabilistic, inductive reasoning are some of the keywords pointing to such additional aspects. Little has been done in view of parallelizing such more complicated (though essential) inference techniques.

Another controverse discussion is lead on the question to what extent the user should control the task-partitioning and thus the parallelism. More generally, how should we program a machine with say 64K processors that operate In an asynchronous way (because of the differences in the various parts a synchronous behavior seems to be unrealistic). There is a whole spectrum of opinions on these programming language aspects that might reflect the different possible user levels, ranging from the software engineer to the casual user.

Given so many questions about the nature of the task of inferencing in general and its inherent parallelism in particular, it is no surprise that we still lack a convincing proposal for an architecture of a parallel inference machine. What should be its topology, the power of its processors, the mechanisms of its communications, synchronization, and load-balancing, these are some of the questions that are currently studied in many laboratories around the world.

In this situation it was thought that a panel discussion would provide the appropriate forum to serve a number of functions. It might give a feel for the relevance of each of the controverse discussions mentioned above and the different standpoints taken in them. At best, it might even help to provide some answers to the questions of concern. In any case it will inform the AI-community that much more is going on in this promising area than one might think from the relatively sparse publications, which is typical for any field during its initial phase of experimentation.