Subject: Evaluation of Ph.D. thesis submitted by Zbynek Falt

The Ph.D. Thesis of Mr. Falt deals with the important and timely topic of parallel software for data-intensive applications on current and future parallel architectures. In particular in his thesis he proposes a new higher-level language, called Bobolang for the specification of execution plans for parallel data processing applications, several enhancements and optimizations of the associated Bobox parallelization framework developed at the Charles University in Prague, and several novel parallel data processing algorithms.

A first major contribution is the design and implementation of a new task scheduler for the Bobox framework that deals with the scheduling of execution plans for pipelined task-parallel data processing applications. Although this scheduler shares some similarities with related schedulers (in particular Intel's TBB scheduler), its main innovative aspect is the fact that it is NUMA-aware. The author has demonstrated that by carefully considering the location of data in the memory hierarchy of NUMA architectures, the performance of execution plans can be significantly improved. In addition to the NUMA-aware scheduling strategy the author has also enhanced the API for stream processing used within Bobox and has implemented a new memory allocation strategy that aims at reusing already allocated virtual memory as much as possible.

A second major contribution is the development of Bobolang, a new language for the specification of execution plans for parallel data processing applications, as well as associated compilation and parallelization techniques. Bobolang allows the specification of operators and execution plans at a high-level of abstraction. Specific operators for describing parallelization have been designed in order to enable semi-automatic extraction of intra-operator parallelism. An outline of the implementation of Bobolang language is provided along with experimental results.

As a third contribution, the thesis presents several novel parallel algorithms for data intensive computations. This includes new algorithms for sort and merge-join, which have been designed specifically for the Bobox environment with the goal of achieving high scalability. In addition, a special focus was put on designing these algorithms in such a way that they attempt to maintain load balance also in the case of skewed data. Finally, a parallel runtime support system for SPARQL has been implemented and the provided experimental results using the sp2bench benchmark indicate that it achieves very good performance comparable to the fastest available in-memory SPARQL engines.

In general the presented thesis is very well written, methodologically sound, and contains several novel and significant scientific contributions with respect to the state of the art in this field, which also has been well reviewed in this thesis. All the proposed innovations have been implemented and experimentally evaluated within the Bobox framework.
The author has already published parts of his work in peer-reviewed publications at top international conferences and journals and has clearly demonstrated his ability to undertake independent and innovative scientific research.

For all these reasons I strongly suggest to accept the presented work as a Ph.D. thesis with the highest possible mark.

Siegfried Benkner
Faculty of Computer Science
University of Vienna