

Systems for high performance computing are getting increasingly complex. On the one hand, the number of processors is increasing. On the other hand, the individual processors are getting more and more powerful. In recent years, the latter is to a large extent achieved by increasing the number of cores per processor. Unfortunately, scientific applications often fail to fully utilize the available computational performance. Therefore, performance analysis tools that help to localize and fix performance problems are indispensable. Large scale systems for high performance computing typically consist of multiple compute nodes that are connected via network. Performance analysis tools that analyze performance problems that arise from using multiple nodes are readily available. However, the increasing number of cores per processor that can be observed within the last decade represents a major change in the node architecture. Therefore, this work concentrates on the analysis of the node performance.

The goal of this thesis is to improve the understanding of the achieved application performance on existing hardware. It can be observed that the scaling of parallel applications on multi-core processors differs significantly from the scaling on multiple processors. Therefore, the properties of shared resources in contemporary multi-core processors as well as remote accesses in multi-processor systems are investigated and their respective impact on the application performance is analyzed. As a first step, a comprehensive suite of highly optimized micro-benchmarks is developed. These benchmarks are able to determine the performance of memory accesses depending on the location and coherence state of the data. They are used to perform an in-depth analysis of the characteristics of memory accesses in contemporary multiprocessor systems, which identifies potential bottlenecks. However, in order to localize performance problems, it also has to be determined to which extend the application performance is limited by certain resources. Therefore, a methodology to derive metrics for the utilization of individual components in the memory hierarchy as well as waiting times caused by memory accesses is developed in the second step. The approach is based on hardware performance counters, which record the number of certain hardware events. The developed micro-benchmarks are used to selectively stress individual components, which can be used to identify the events that provide a reasonable assessment for the utilization of the respective component and the amount of time that is spent waiting for memory accesses to complete. Finally, the knowledge gained from this process is used to implement a visualization of memory related performance issues in existing performance analysis tools. The results of the micro-benchmarks reveal that the increasing number of cores per processor and the usage of multiple processors per node leads to complex systems with vastly different performance characteristics of memory accesses depending on the location of the accessed data. Furthermore, it can be observed that the aggregated throughput of shared resources in multi-core processors does not necessarily scale linearly with the number of cores that access them concurrently, which limits the scalability of parallel applications. It is shown that the proposed methodology for the identification of meaningful hardware performance counters yields useful metrics for the localization of memory related performance limitations.