

Effective & Adaptive Parallelism

SECTION A - PROJECT SUMMARY

With the advent of new fabrication technologies, integrated circuit designs can utilize more than hundred million gates. The large increase in transistor count has significantly increased device complexities allowing to implement entire complex systems on single chips (SOC) and even massively parallel systems could fit on a single chip in the next chip generations. Despite the increasing device complexities cost is still an important issue in most digital designs, in particular for targeting low-power. In fact only few designs really may use as many gates as technologically possible. Most markets for digital components (e.g. mobile devices) impose strict limitations on the cost and the power dissipation of a chip. Gates and switching activity therefore have to be used judiciously. Performance often needs to be optimised within given transistor and power budgets.

Intellectual merit of the proposed activity. We address this system wide performance optimisation problem for fixed transistor and power budgets through the effective utilization of parallelism. Parallelism can be (and for high-performance it has to be) exploited at many different granularities (e.g. gate level, instruction level, thread level). The effects and dependencies of parallelism at different granularity levels are not yet well understood at all levels and have not yet been considered in an integrated approach, but their integrated consideration is essential to system wide design optimisation of any digital system. The effectiveness of parallelism at each granularity level may vary independently based on various properties of the application. We propose the simultaneous consideration of parallelism at different granularities in a system wide approach. For fixed system cost the excessive gate count for increased parallelism at one granularity level can be made up with reduced gate count of reduced parallelism at other granularities. We develop parametric models that allow the independent scaling of implementations at each granularity level and the customisation of various implementation options in a general framework. Implementations from our framework are evaluated with technology-independent hardware models for cost, performance and power dissipation. It is a major objective of our work to determine the parameter dependencies in our system models for fixed system cost or power dissipation. We consider parametric optimisation of the performance among system organizations of the same cost or power. The main variation in system architecture we consider is the trading of parallelism between different granularities. We focus our consideration on selected basic classes of (parallel) systems. For ASIC designs the system organization is statically adopted for an application (class). We develop prototype implementations on reconfigurable hardware and provide parametric module implementations and parametric system generators with open sources. For reconfigurable hardware we additionally consider the run-time adaptation of the system organization to dynamically changing application properties for improved performance.

Broader impacts of the proposed activity. Performance optimisation for fixed system cost or power is a problem that is often arising in practical contexts. Conventionally, time consuming exploration of design options and numerous design iterations are necessary for this purpose. Our constructive approach allows to simplify design exploration and to significantly accelerate the design process especially for untypical application domains. In particular the non-expert designer is assisted with optimised system generation. Parallel optimisation at different granularities and their trade-off has not yet been considered in an integrated approach. The fundamental analysis and the general framework of our work provides and enhances fundamental understanding of parallelism at different granularities and their interaction. Our development of prototypes, module libraries and system generator tools extend the collection of design support with open sources that are accessible through the WEB. In particular we plan the creation of a free WEB-service for "Application Specific System-on-Demand Designs". Among others the development and application of accurate technology-independent hardware models in our work demonstrates their applicability for complex digital systems and prepares for their broader use in future scientific system evaluations. The industrial impact of this project is further supported by the attached letter of reference from AMD.

Integration of Research and Education. Our Research developments allow the integrated consideration of large families of (parallel) computer organizations and the detailed study of many effects in their design. Fundamental issues in high-performance and low-power design can be discussed in a quantitative, integrated and system-wide approach. Our quantitative analysis considers technology-independent hardware models for performance, cost and power dissipation that go far beyond existing quantitative studies (e.g. Hennessy/Patterson) in generality, amount of detail and accuracy. Interestingly, instances for applied gate level parallelism (like floating-point and optimised arithmetic) are typically discussed separately (in an appendix in HP) in all previous books on computer architecture and rarely contain detailed cost and power evaluations. The integration of our research results into the curriculum will enhance the learning experience on digital systems at several levels. Students from these classes can vice versa contribute in the research part of this project not only for their PhD studies. The installation of classroom design competitions, integrated research components in graduate and undergraduate courses and undergraduate research projects will further foster the tight coupling of the research and education components.