Towards High Performance and Efficiency of Distributed Heterogeneous Systems

Sam Skalicky, Sonia Lopez and Marcin Lukowiak

Department of Computer Engineering, Rochester Institute of Technology, Rochester, NY.

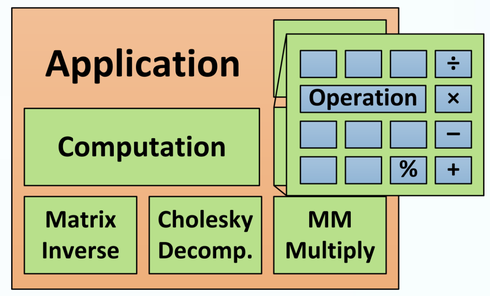
HiPer High Performance Research Group

Overview

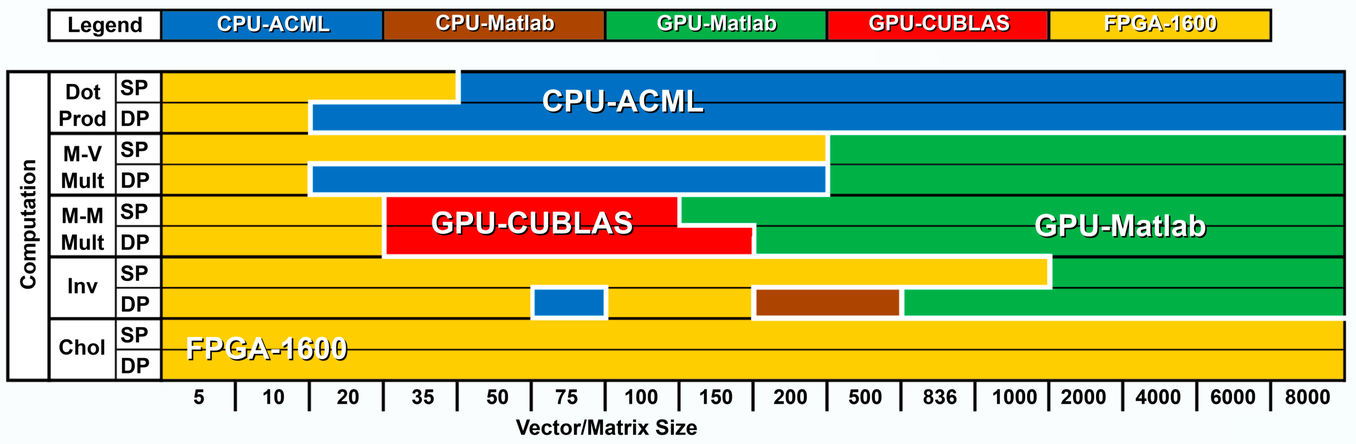
- Goal: Ease the design and performance tuning of applications in heterogeneous systems
- Implementation problems[1]:
  - Diverse hardware platforms: CPU, GPU, FPGA
  - Profiling & Benchmarking
  - Partitioning & Mapping (task granularity)
  - Scheduling & Synchronization
  - Performance Evaluation
- Proposed Solution: Model-based framework
  - Accounts for various computations and input data sizes
  - Understands different capabilities of hardware platforms
  - Evaluates various implementations for each computation
  - Enable regular use of compute-intensive applications by achieving performance requirements
  - Heterogeneous systems provide variety of different computational capabilities[2]

Contributions

- High-level graph-based models for estimating computation execution time on any hardware platform without implementation[4,5]
- Framework is a tool to transform a single-threaded application into parallel implementation that uses various hardware platforms
- Our Target Applications: those with diverse workloads that incorporate large amounts of data such as: medical diagnosis, weather prediction, and stock & securities market analysis
- Decompose applications into core computations (often linear algebra)



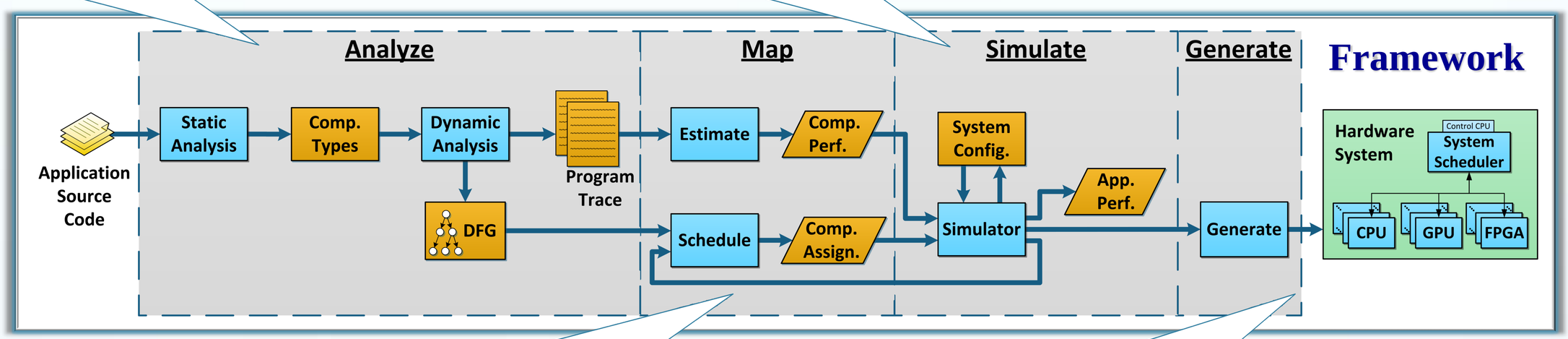
Application decomposition



Design Space Chart showing the highest performing platforms and implementations. The different regions represent the platform with the best performance at any computation and data size for single precision (SP) and double precision (DP) floating point.

Analyze – Extract application information
- Static analysis – identify computations
- Dynamic analysis – create dataflow graph (DFG)

Simulate – Estimation application performance
- Determine configuration quantity & type of platforms in the system
- Apply scheduling policy to determine application performance

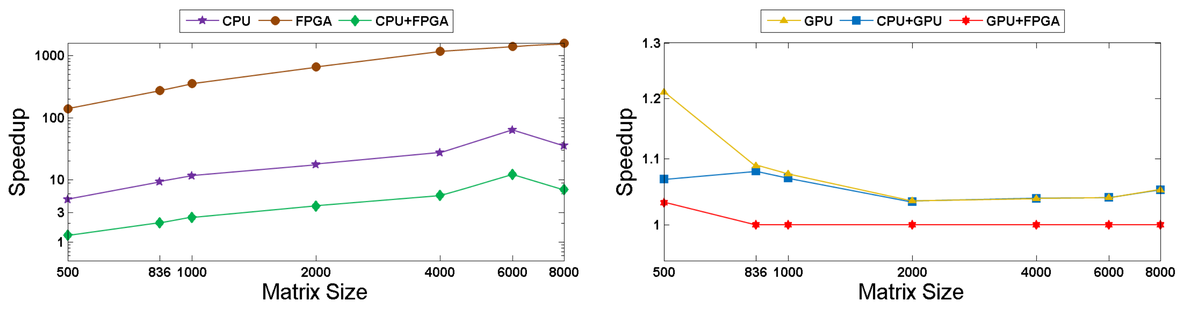


Map – Estimate computation performance, assignments
- Estimate – performance of various platforms
- Schedule – determine computation-to-hardware assignments & choose scheduling policy

Generate –Construct system specification
- Configuration – quantity and type of platforms
- Communication interfaces – network topology
- System Control & Scheduler – manage execution

Results

- Analyzed medical imaging application[3]
- Simulated 7 possible system configurations:
  - Single platform: CPU, GPU, or FPGA
  - Two platform: CPU+GPU, CPU+FPGA, GPU+FPGA
  - Three platform: CPU+GPU+FPGA
- Performance results:[6]
  - Three platform system performed best
  - Results split into systems with GPU (right) and without GPU (left)
  - GPU platform provided largest benefit, FPGA 2nd, CPU 3rd overall
  - Combining CPU+GPU+FPGA achieves 62x, 2x, and 1605x speedups compared to single CPU, GPU, or FPGA platform systems



Performance of each system normalized to performance of CPU+GPU+FPGA system

Conclusion

- Investigated systems comprised of hardware platforms from commercially available off-the-shelf components.
- Results are applicable to future on-chip heterogeneous systems.
- Due to increasing system complexity analysis tools will be critical to aid in the software design to take advantage of future hardware.
- The performance of computationally intensive applications can be improved using heterogeneous systems.

References

1. A. Khokhar, V. Prasanna, M. Shaaban, and C.L. Wang, *"Heterogeneous Computing: Challenges and Opportunities,"* Computer, vol. 26, no. 6, 1993.
2. S. Skalicky, S. Lopez, M. Lukowiak, J. Letendre, D. Gasser, *"Linear Algebra Computations in Heterogeneous Systems,"* IEEE Intl. Conf. on Application-specific Systems, Architectures and Processors, 2013.
3. L. Wang, H. Zhang, K. C. L. Wong, H. Liu, and P. Shi, *"Physiological-model-constrained Noninvasive Reconstruction of Volumetric Myocardial Transmembrane Potentials,"* in IEEE Trans. on Biomedical Engineering, 2010.
4. S. Skalicky, S. Lopez, M. Lukowiak, *"A Scheduling Approach for Performance Estimation on Multiprocessor Architectures,"* The Intl. Conf. for High Performance Computing, Networking, Storage and Analysis, 2014, under review.
5. S. Skalicky, S. Lopez, M. Lukowiak, *"High-Level Graph-Based Methodology for Improving Performance of Pipelined Architectures,"* Intl. Conf. on Field Programmable Logic and Applications, 2014, under review.
6. S. Skalicky, S. Lopez, M. Lukowiak, *"Distributed Execution of Transmural Electrophysiological Imaging with CPU, GPU, and FPGA,"* Intl. Conf. on ReConFigurable Computing and FPGAs, 2013.