Multilevel scheduling of computations in large-scale parallel computing systems

Ahmed Eleliemy and Florina. M. Ciorba (Ph.D. Advisor)
Department of Mathematics and Computer Science, University of Basel, Switzerland

1. What is it about?

- ***Batch level scheduling*** (BLS) ***competitively*** schedules jobs on shared HPC systems.
- ***Application level scheduling*** (ALS) and ***thread level scheduling*** (TLS) ***cooperatively*** schedule computational workloads on a given set of compute resources.

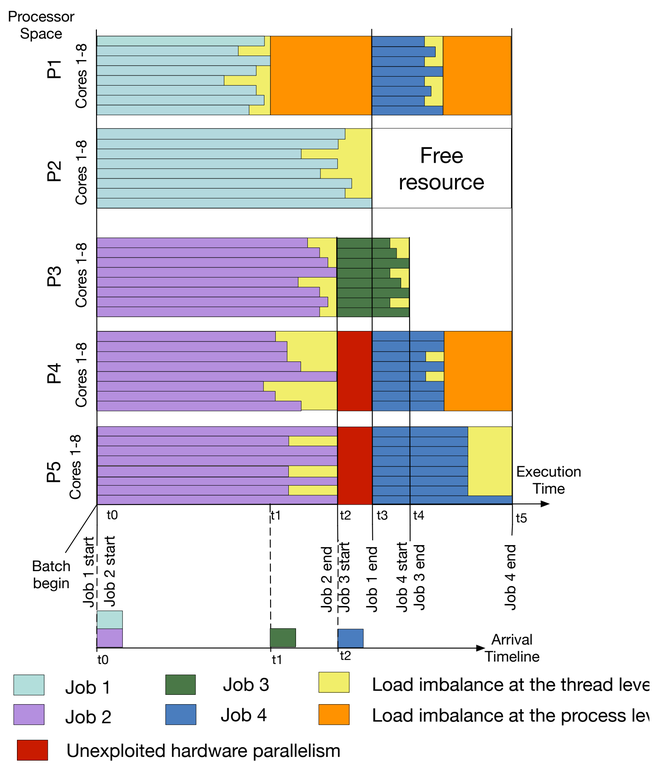
Research Question

Given the competitive aspect in the BLS and the cooperative aspect in the ALS and TLS, how we can enable feedback and live information between these levels of scheduling to enhance the parallel execution of scientific applications on HPC systems?

2. Objectives by Priority

1. Minimize application execution times.
2. Maximize resource utilization.
3. Maximize system throughput.

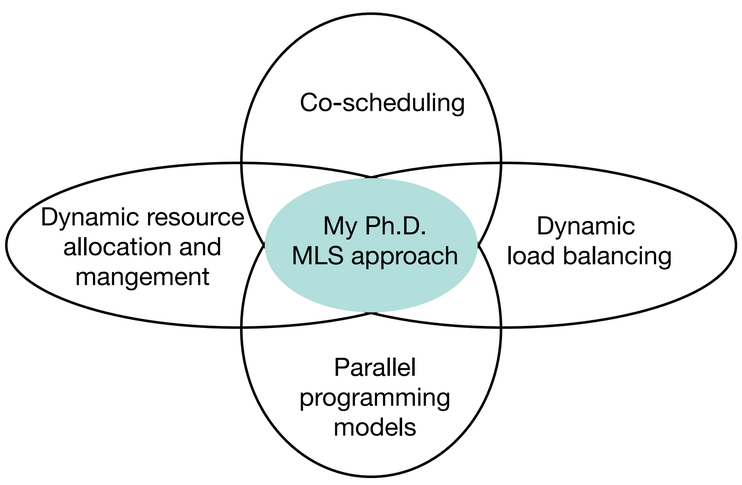
3. State of the Art



The total number of computing resources assigned to a certain job do not change at the runtime

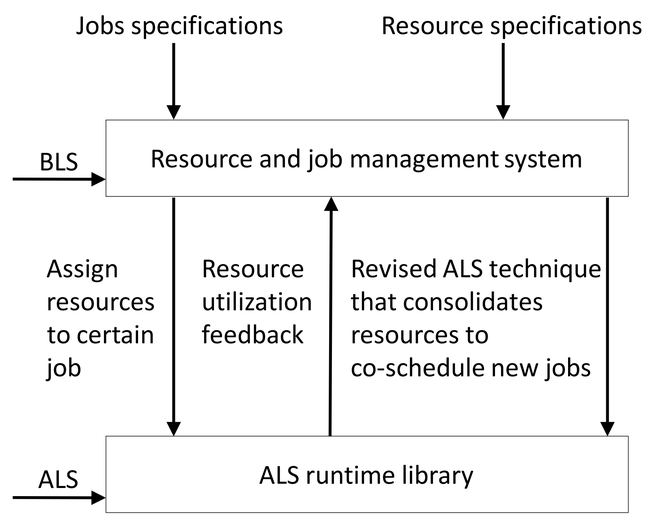
4. Envisioned Strategy

- ***Eliminate*** resource fragmentation ***via dynamic load balancing***.
- ***Exploit*** unavoidable resource fragmentation ***via resource consolidation and co-scheduling***.
- The Multilevel scheduling (MLS) approach is the ***intersection*** between the following research areas



4. How?

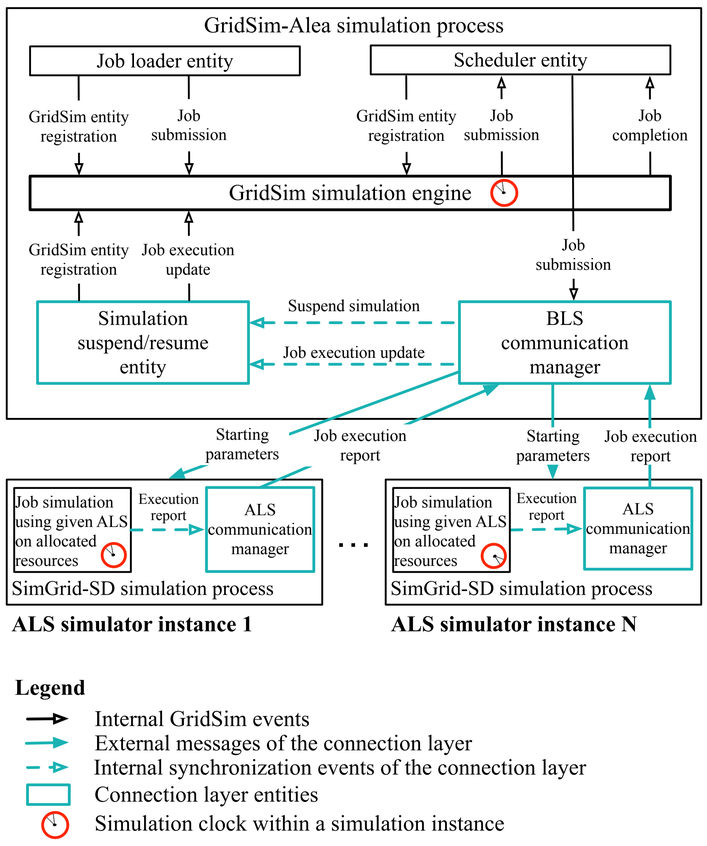
Changing the application scheduling techniques on a certain resource to co-schedule more applications on that resource



6. Accomplishments

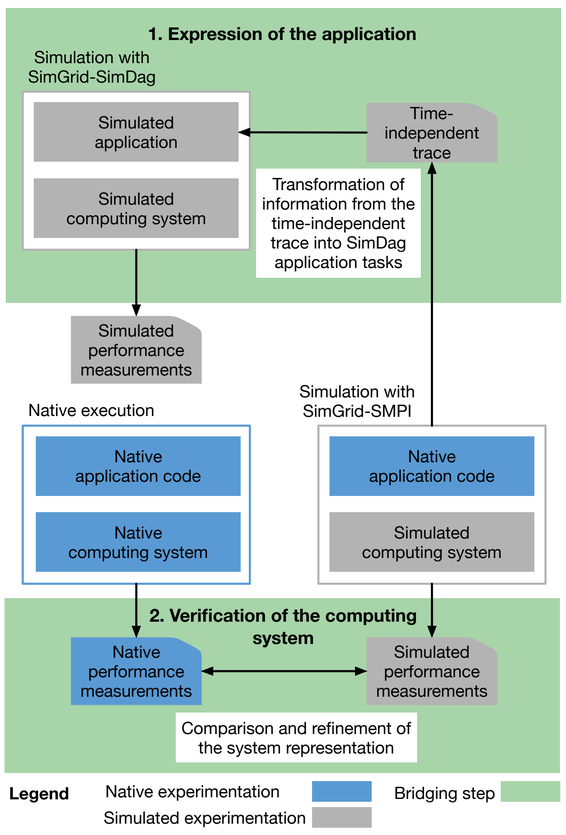
Accomplishment 1:
A two-level simulator for ***batch*** and ***application*** level scheduling [1]

BLS simulator instance



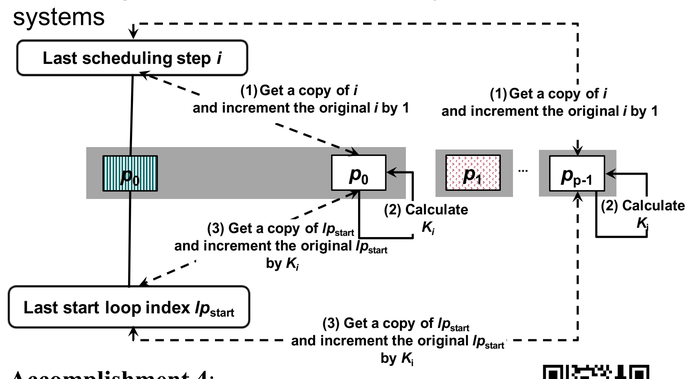
- A generic simulation approach that bridges two existing simulators from batch and application level scheduling
- A study of exploring the relationship between BLS and ALS

Accomplishment 2:
Methodology for bridging native and simulated executions of parallel applications on HPC systems [3]

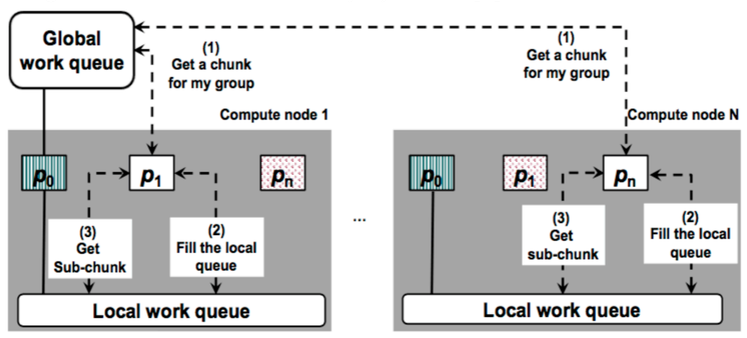


- A method for obtaining high confidence in the results obtained natively and via simulation
- Evaluates the usefulness of using FLOP count vs. time-based measurements to represent the application characteristics in simulation

Accomplishment 3:
A distributed chunk-calculation approach for dynamic [2,4] loop scheduling at ALS on heterogeneous distributed-memory systems



Accomplishment 4:
A Hierarchical distributed chunk-calculation approach for dynamic loop scheduling on large scale distributed memory systems [5]



- Proposal and implementation of a hierarchical version of DLS techniques for distributed-memory systems using the MPI+MPI approach.
- Evaluation of using the MPI+MPI approach in developing hierarchical DLS techniques

7. Next Steps

- Enable sending live information from DLS to the resource and job manager.
- Evaluate BLS+ALS via simulaive and native experiments.
- Extend the approach to connect TLS with BLS+ALS.

8. Selected Ph.D. Publications

[1] Eleliemy, A., Mohammed, A., and Ciorba, F. M. "Exploring the Relation Between Two Levels of Scheduling Using a Novel Simulation Approach", The 16th International Symposium on Parallel and Distributed Computing (ISPDC), 2017.

[2] Eleliemy, A., Mohammed, A., and Ciorba, F. M. "Efficient Generation of Parallel Spin-images Using Dynamic Loop Scheduling", The 8th International Workshop on Multicore and Multithreaded Architectures and Algorithms (M2A2) of the 19th IEEE International Conference for High Performance Computing and Communications (HPCC), 2017.

[3] Mohammed, A., Eleliemy, A., Ciorba, F. M., Kasielke, F., and Banicescu, I. "Experimental Verification and Analysis of Dynamic Loop Scheduling in Scientific Applications", The 17th International Symposium on Parallel and Distributed Computing (ISPDC), 2018.

[4] Eleliemy, A. and Ciorba, F. M. "Dynamic Loop Scheduling Using MPI Passive-Target Remote Memory Access", The 27th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP), 2019.

[5] Eleliemy, A. and Ciorba, F. M. "Hierarchical Dynamic Loop Scheduling on Distributed-Memory Systems Using an MPI+MPI Approach", The 20th IEEE International Workshop on Parallel and Distributed Scientific and Engineering Computing (PDSEC), 2019.