MLS: Multilevel Scheduling in Large Scale High Performance Computers

Florina M. Ciorba (Principle Investigator), Ahmed Eleliemy, and Jonas H. Müller Korndörfer
Department of Mathematics and Computer Science, University of Basel, Switzerland
{firstname.lastname}@unibas.ch

Duration: September 1, 2017 – August 31, 2020
hpc.dmi.unibas.ch/HPC/MLS.html

Swiss National Science Foundation

1. Research Problem

Research Question

Given massive parallelism, at multiple levels, and of diverse forms and granularities, how can it be ***exposed***, ***expressed***, and ***exploited*** such that execution times are reduced, performance targets are achieved, and acceptable efficiency is maintained?

Methodology

Multilevel scheduling (MLS) extends and bridges the most successful batch (jobs scheduled on a system), application (application processes scheduled on allocated nodes/processors/cores), and thread (application threads scheduled on the allocated cores) scheduling approaches beyond a single or a couple levels of parallelism (scaling across) and beyond their current scale (scaling out).

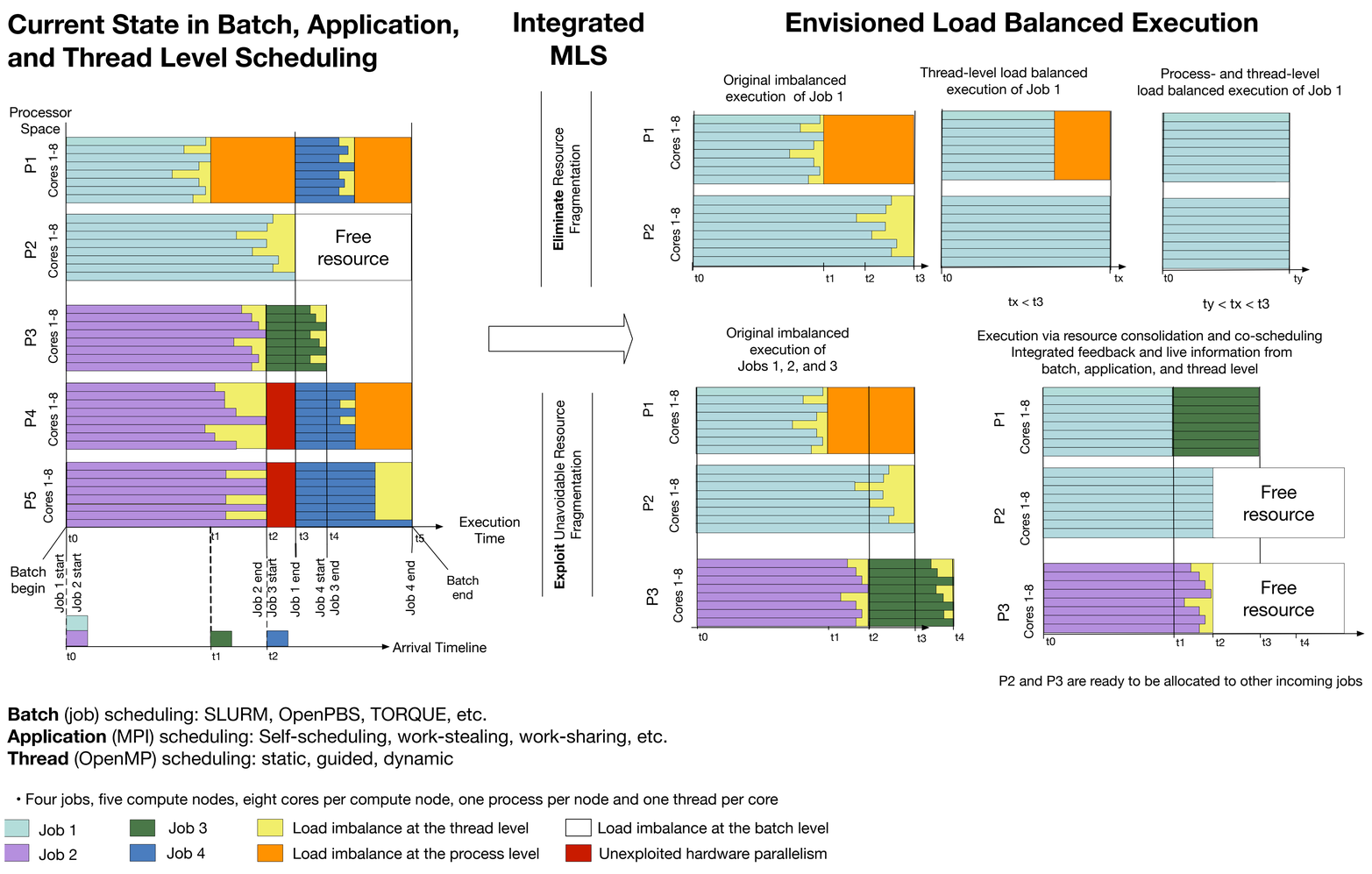
Envisioned Outcomes

A prototype multilevel scheduling solution that integrates live feedback information from three, currently disjoint, scheduling levels: batch, application, and thread.

2. Objectives

1. Leverage all available parallelism **within** each single hardware parallelism level and **across** the three hardware parallelism levels (system, node, core).
2. Address hardware heterogeneity.
3. Achieve robustness against perturbations (including variations and failures) while minimizing execution time, maintaining acceptable efficiency, and maximizing resource utilization.

3. Exposing, Expressing, and Exploiting Multilevel Parallelism

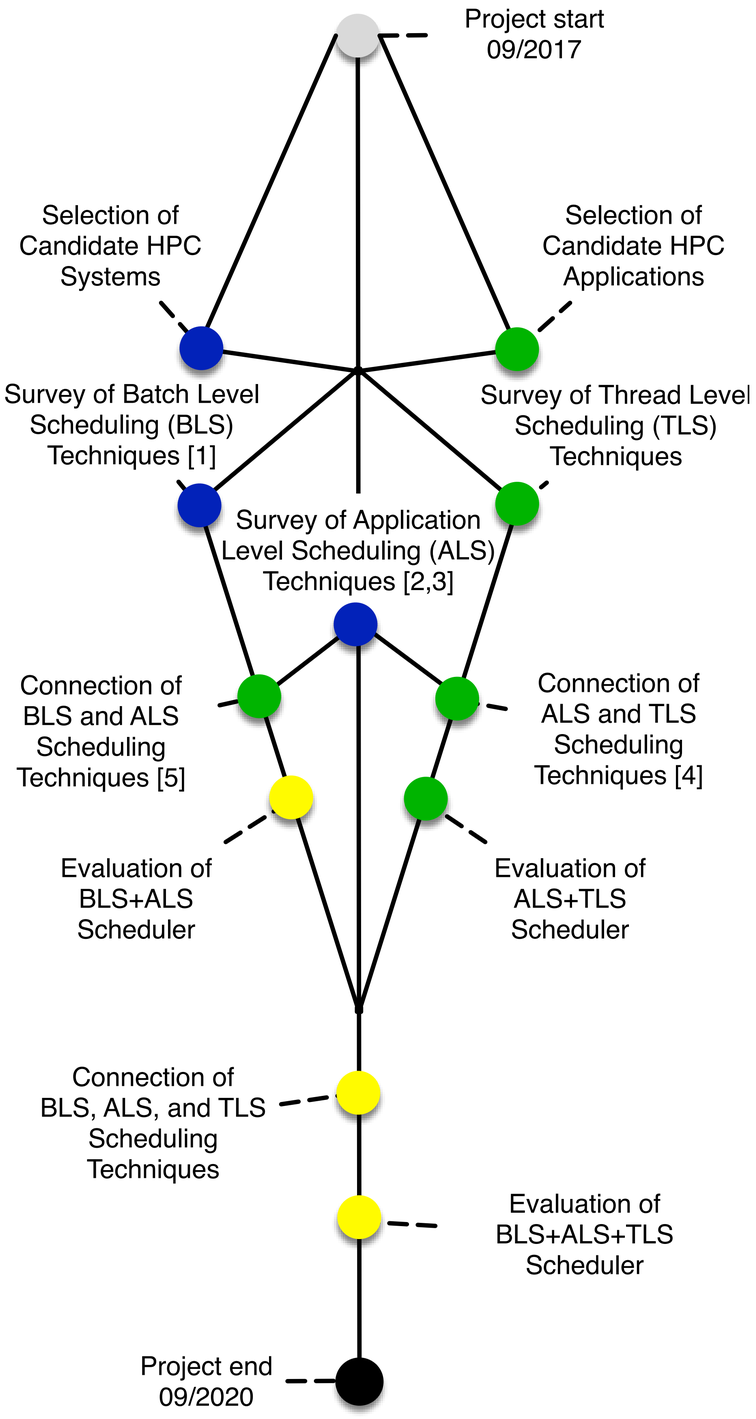


Batch (job) scheduling: SLURM, OpenPBS, TORQUE, etc.
Application (MPI) scheduling: Self-scheduling, work-stealing, work-sharing, etc.
Thread (OpenMP) scheduling: static, guided, dynamic

• Four jobs, five compute nodes, eight cores per compute node, one process per node and one thread per core

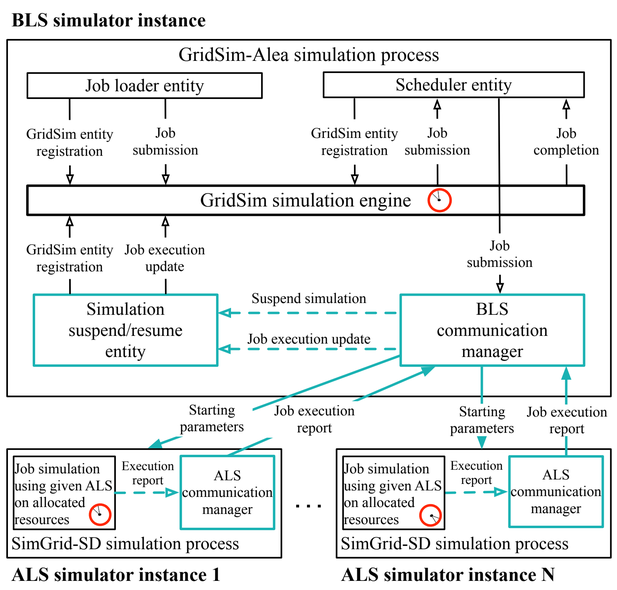
Job 1 · Job 2 · Job 3 · Job 4 · Load imbalance at the thread level · Load imbalance at the process level · Load imbalance at the batch level · Unexploited hardware parallelism

4. Project Status

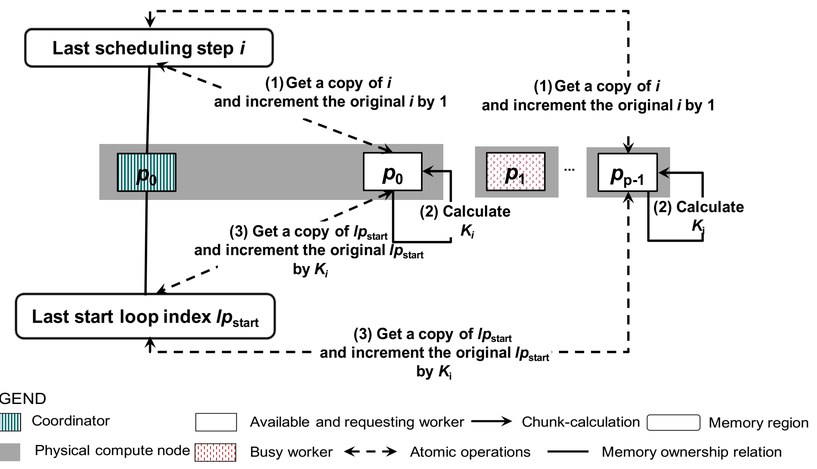


5. Selected Achievements

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



Achievement 2: A distributed chunk-calculation approach for dynamic loop scheduling at **application** level scheduling on heterogeneous distributed-memory systems [5,7]



Achievement 1:

Introduced **a generic simulation approach** that bridges two existing simulators from batch and application scheduling levels to explore the relationship between batch and application level scheduling.

Achievement 2:

Introduced a distributed chunk-calculation approach that is **recommended** for developing DLS techniques on **heterogeneous distributed-memory systems** for application level scheduling.

6. Project Publications

[1] Eleliemy, A. and Ciorba, F. M. "*Hierarchical Dynamic Loop Scheduling on Distributed-Memory Systems Using an MPI+MPI Approach*", In Proceedings of the 20th IEEE International Workshop on Parallel and Distributed Scientific and Engineering Computing (PDSEC), 2019.
[2] Eleliemy, A. and Ciorba, F. M. "*Dynamic Loop Scheduling Using MPI Passive-Target Remote Memory Access*", In Proceedings of the 27th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP), 2019.
[3] Mohammed, A., Ciorba, F. M., Cavelan, A., Cabezon, R., and Banicescu, I. "*Identifying Performance Challenges in Smoothed Particle Hydrodynamics Simulations*", Poster at the Platform for Advanced Scientific Computing Conference (PASC), 2019.
[4] Eleliemy, A. and Ciorba, F. M. "*Dynamic Loop Scheduling Using the MPI Passive-Target Remote Memory Access Model*", Poster at the Platform for Advanced Scientific Computing Conference (PASC), 2018.
[5] Mohammed, A., Eleliemy, A., Ciorba, F. M., Kasielke, F., and Banicescu, I. "*Experimental Verification and Analysis of Dynamic Loop Scheduling in Scientific Applications*", In Proceedings of the 17th International Symposium on Parallel and Distributed Computing (ISPDC), 2018.
[6] Mohammed, A. and Ciorba, F. M. "*SiL: An Approach for Adjusting Applications to Heterogeneous Systems Under Perturbations*", In Proceedings of the International Workshop on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Platforms (HeteroPar) of the 24th International European Conference on Parallel and Distributed Computing (Euro-Par), 2018.
[7] Eleliemy, A., Mohammed, A., and Ciorba, F. M. "*Exploring the Relation Between Two Levels of Scheduling Using a Novel Simulation Approach*", In Proceedings of the 16th International Symposium on Parallel and Distributed Computing (ISPDC), 2017.
[8] Eleliemy, A., Mohammed, A., and Ciorba, F. M. "*Efficient Generation of Parallel Spin-images Using Dynamic Loop Scheduling*", In Proceedings of 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.
[9] Mohammed, A., Eleliemy, A., and Ciorba, F. M. "*A Methodology for Bridging the Native and Simulated Executions of Parallel Applications*", Poster at the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2017.