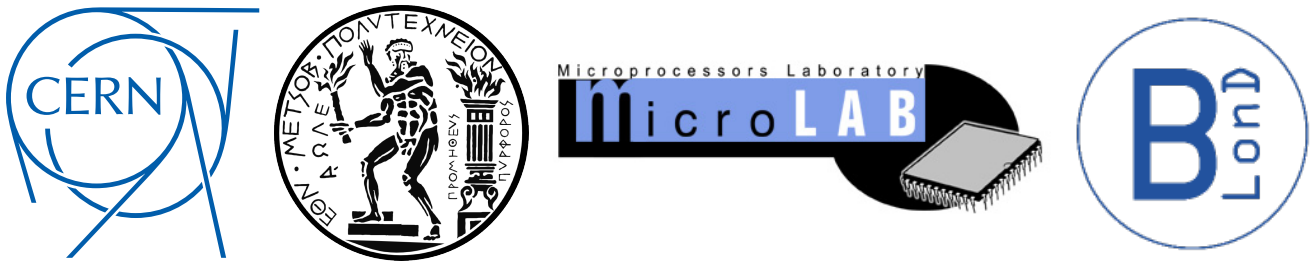
BLonD-MPI: Distributed Longitudinal Beam Dynamics Simulations

Konstantinos Iliakis

PhD Candidate
CERN, CH - NTUA, GR
konstantinos.iliakis@cern.ch

Supervisors:

Dr. Helga Timko, CERN
Dr. Sotirios Xydis, NTUA
Dr. Dimitrios Soudris, NTUA



February 21, 2019

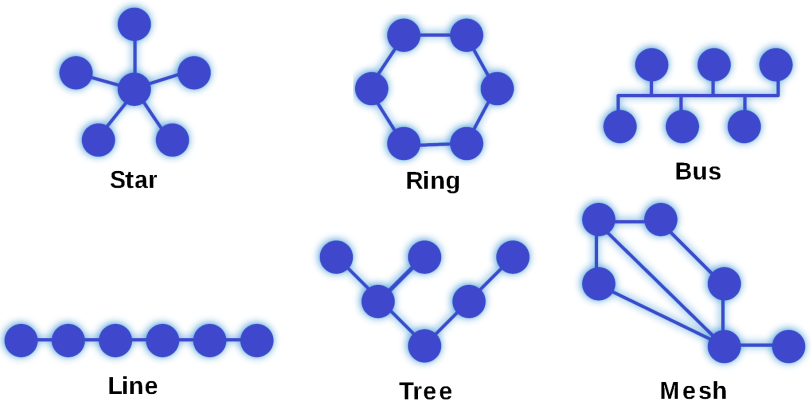
Distributed Computing

Distributed System

- Network of computers **exchanging messages**.
- Perform operations **collectively**.

MPI

- Message Passing **Interface**.
- A **standard** for inter-process communication.
- Various implementations: MPICH, OpenMPI, Intel MPI ...



Motivation

Why we *need* BLonD-MPI?

- Horizontal vs vertical scaling.
- BLonD has been shown to be memory bounded.
- Continuous increase in problem sizes.

Scale Up- Vertical Scaling

Scale Out- Horizontal Scaling

High-level Implementation

Initialization

- All workers (or tasks) execute the script.
- Each worker assigned a subset of the beam.

Main loop

- Each worker tracks its own subset.
- Reduction to generate the global profile.
- Poorly scalable tasks executed by all workers.

Finalization

- A master worker gathers all data back.
- All other workers exit.

Optimizations

Minimize Communication

- Profile casted to 32-bit integer.
- Poorly scalable tasks executed by all workers.

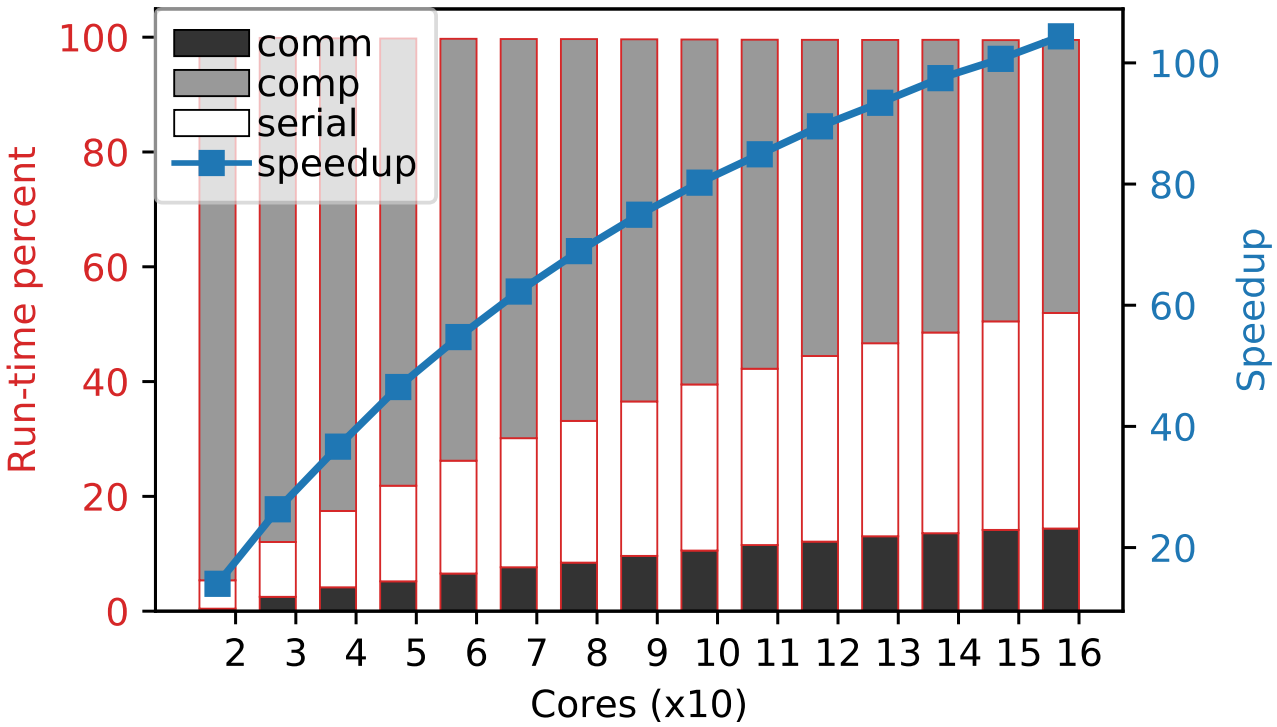
Minimize Serial regions

- All serial regions are parallelised and implemented in C.
- Packed FFTs when multiple induced voltage objects.

Minimize Synchronization

- Only synchronization point: the profile reduction.

Experimental Evaluation



- 72 bunches/ 4Mppb.
- 43K turns.
- BLonD run-time (single core): 2 days.
- BLonD-MPI (8 nodes): 30 minutes.

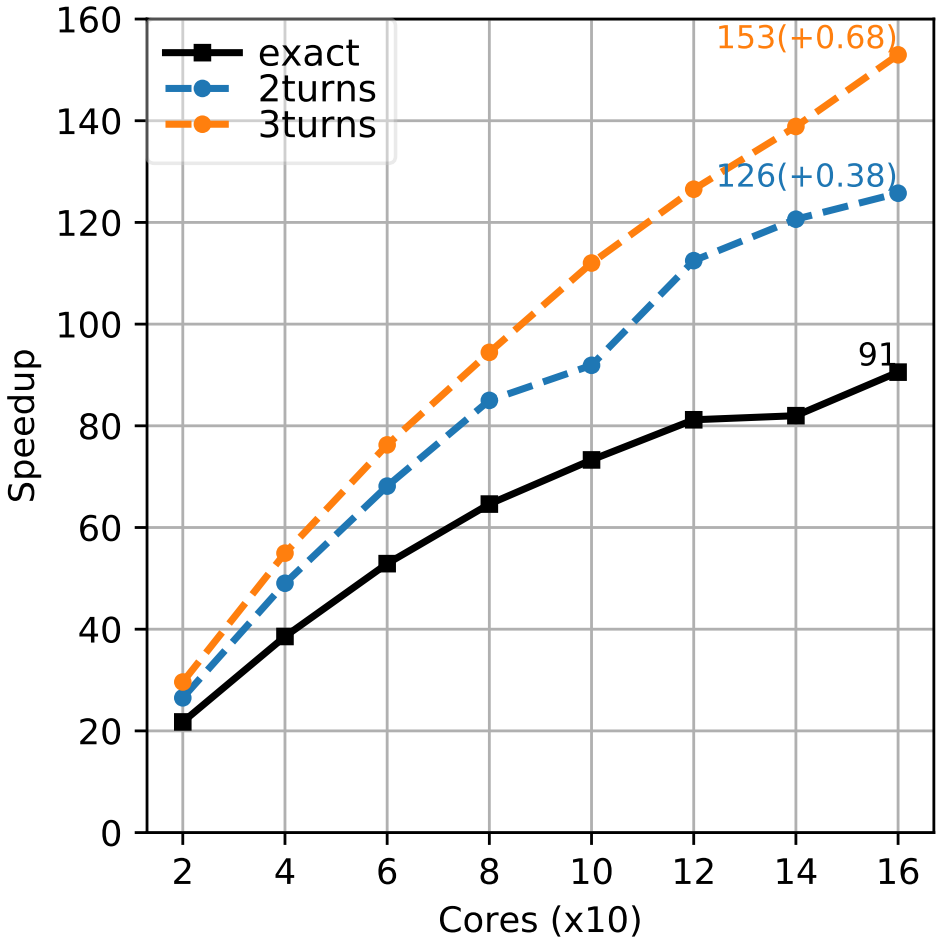
Approximate Computing

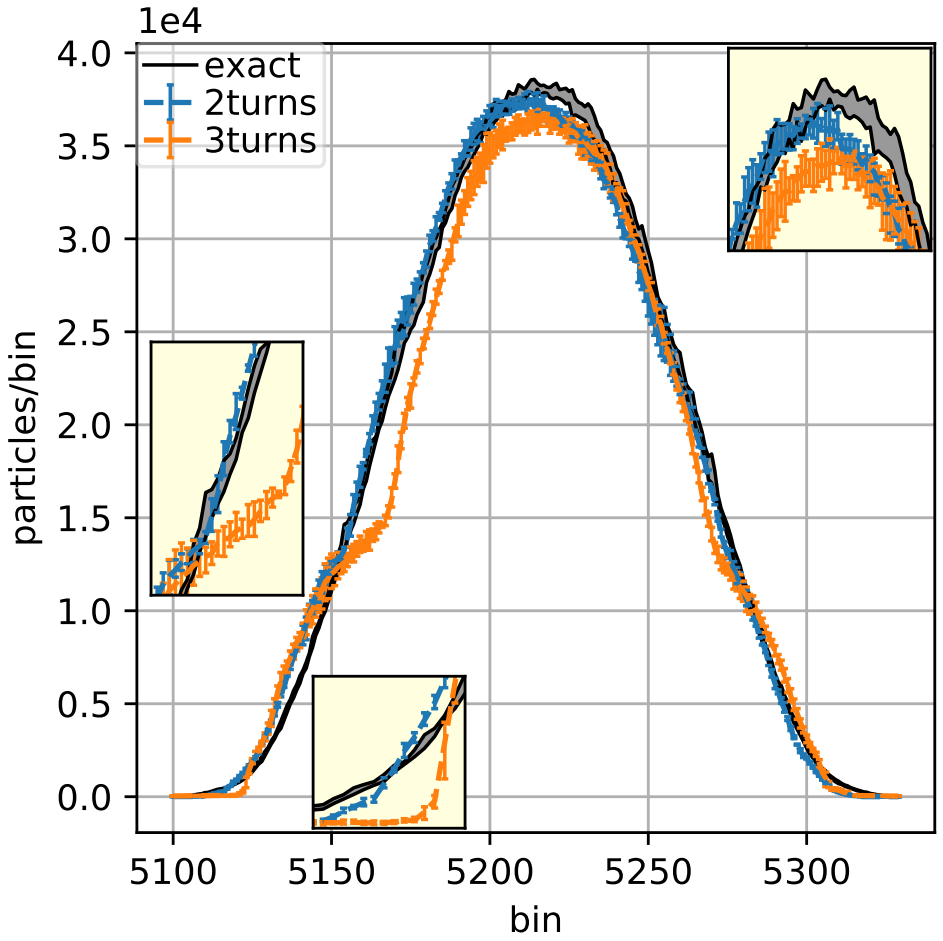
Note that:

- Trade-off accuracy for performance and scalability.
- Useful in the early stage of the design space exploration.
- For advanced users.
- Optional.

Approximate Computing: Method 1

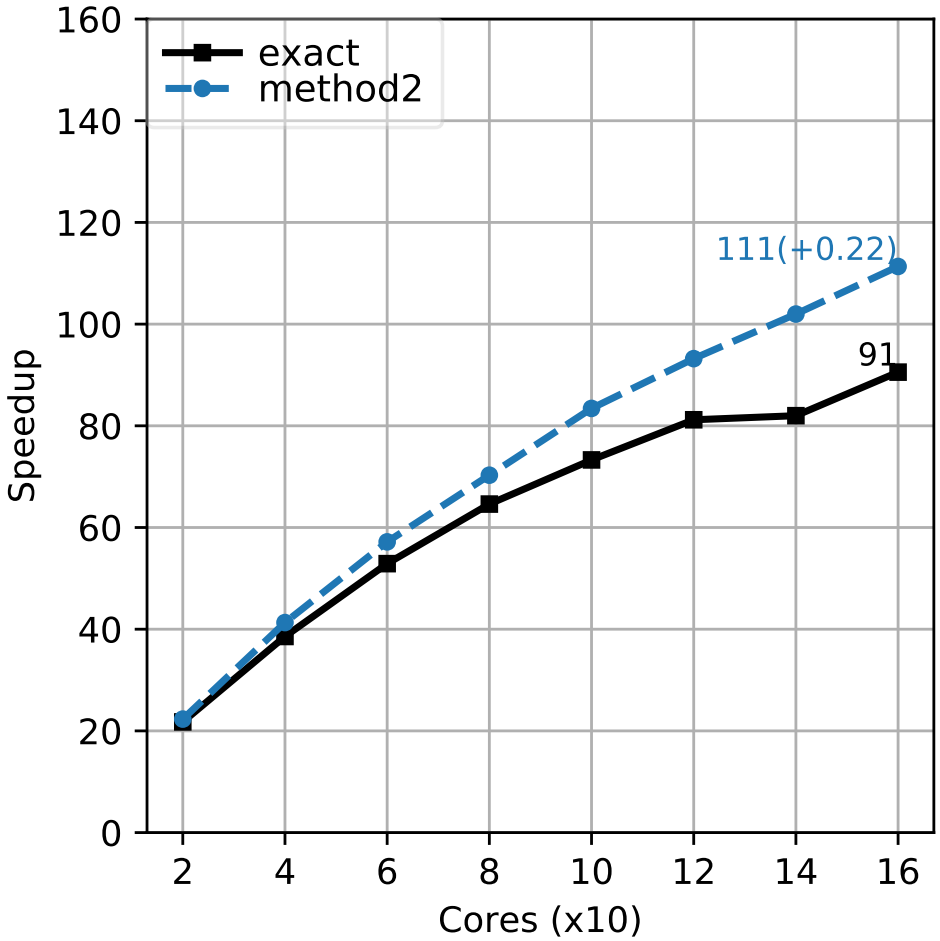
- Assumption: Beam profile changes slightly between consecutive turns.

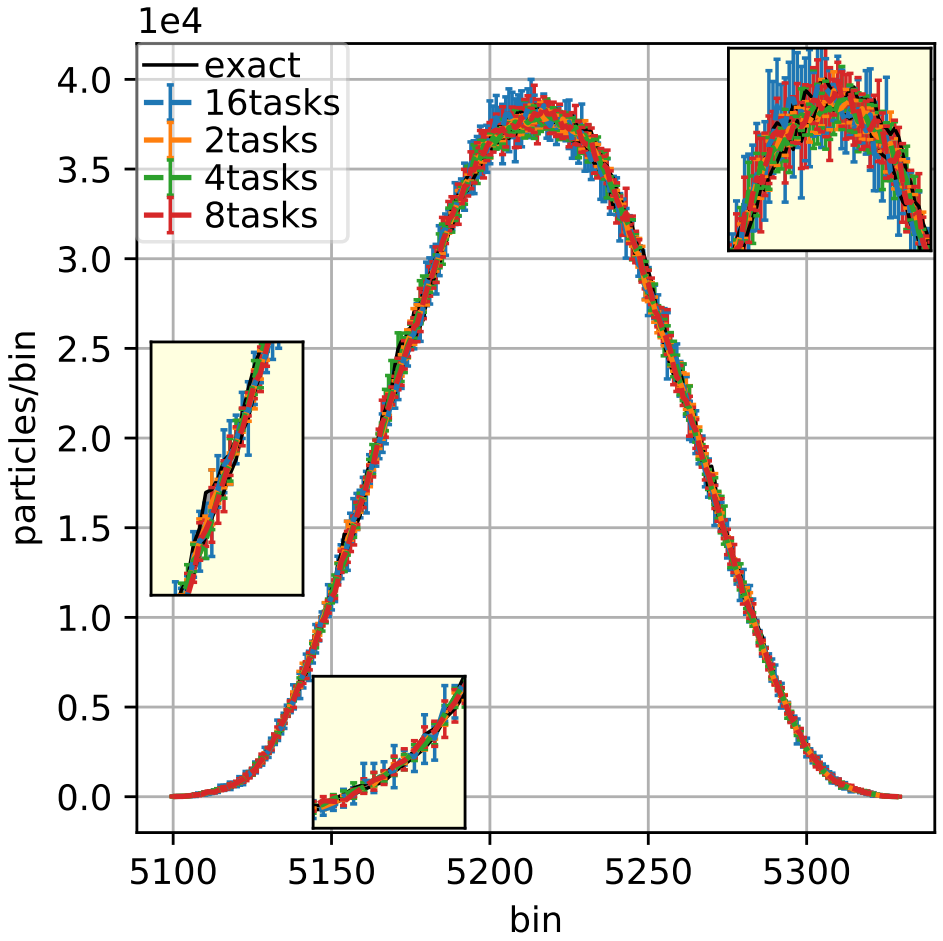




Approximate Computing: Method 2

- Assumption: Every worker is assigned a representative subset of the whole distribution.





CERN HPCBatch Cluster

Practical Info

- Access (members of BE): e-mail Giovanni Rumolo and subscribe to the service-hpc-be e-group.
- 239 Intel nodes $\approx$ 4600 cores.
- All major MPI implementations pre-installed.
- Useful links: Cluster knowledge base and SLURM docs.

Getting started with BLonD-MPI

- Step-by-step instructions on the BLonD-MPI repository page.
- Few (4-5 lines of code) modifications needed in the main file.

Conclusion & Future Work

Conclusions

- BLonD-MPI
  - Uses the power of distributed computing.
  - Reduces the run-time by two orders of magnitude (a year in three days).
  - Enables new studies that were prohibitive in the past.

Future Work

- Merge project with BLonD.
- Run-time manager to bound the approximation error.
- MPI over GPUs (CUDA/ Thrust/ OpenACC).

Q&A

Thanks to my supervisors and Markus Schwarz.