

# Dynamic Load Balancing for BLonD-MPI

Konstantinos Iliakis

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

Supervisors:

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



May 10, 2019

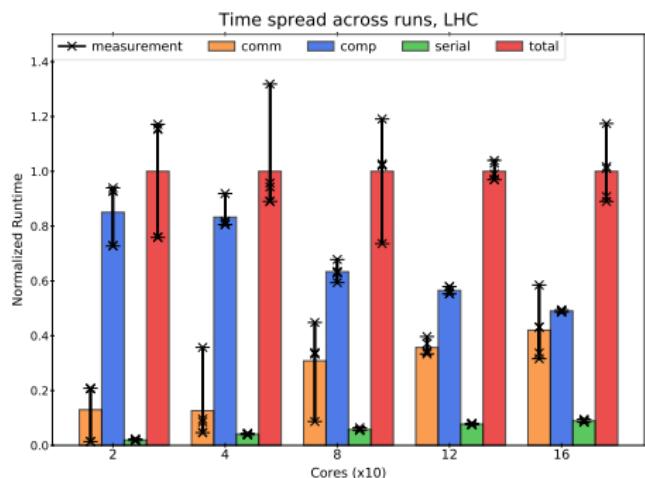
# Motivation

## The Problem

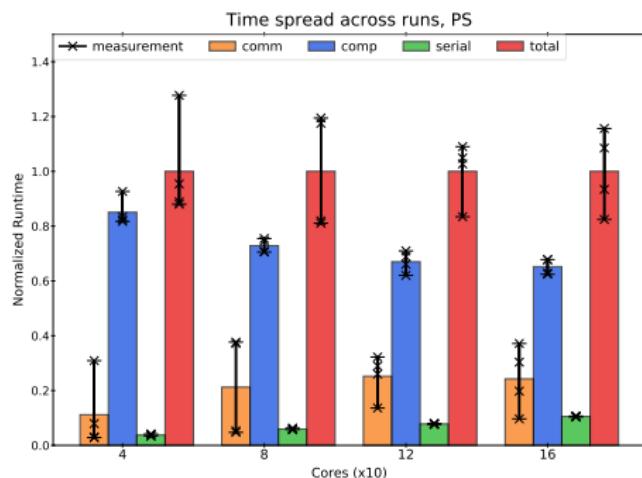
- Node allocations more favourable than others.
- Workers faster than others.
- Impossible or very hard to control the node allocation/ speed of workers.
- Result: Huge spread in run-time across workers and runs.

## Spread across runs(I)

Testcase: LHC

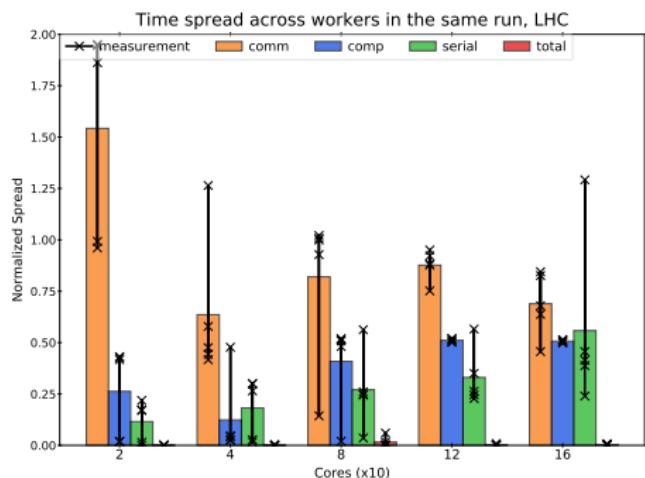


Testcase: PS

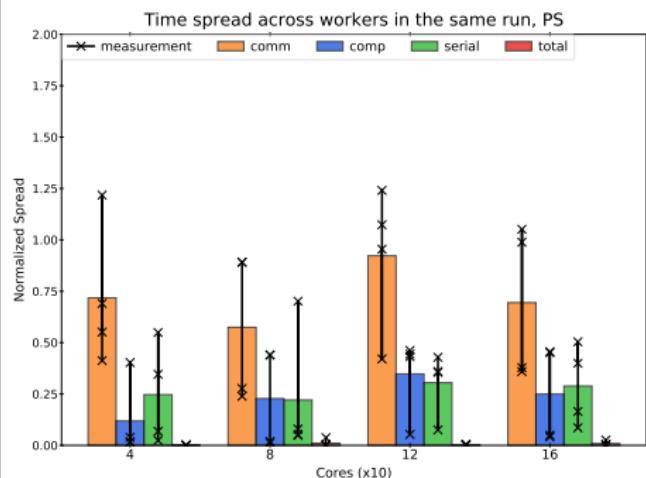


# Spread across workers(I)

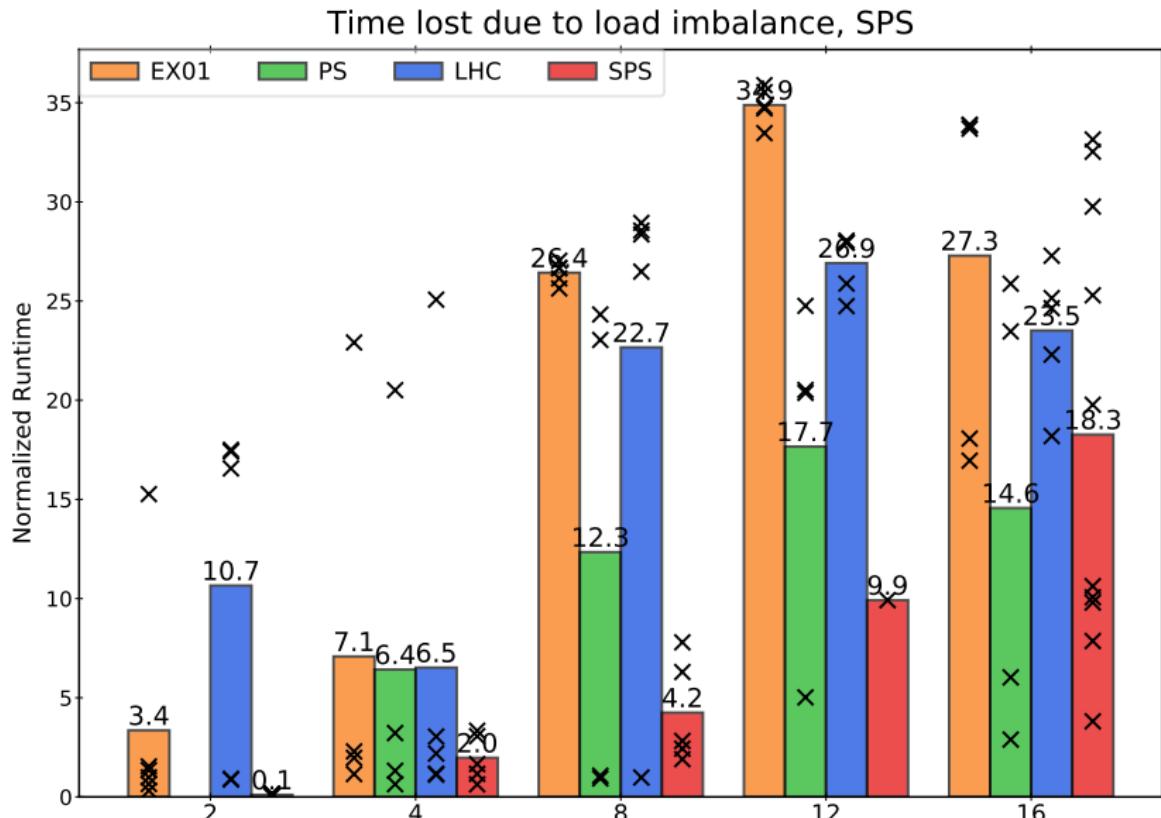
Testcase: LHC



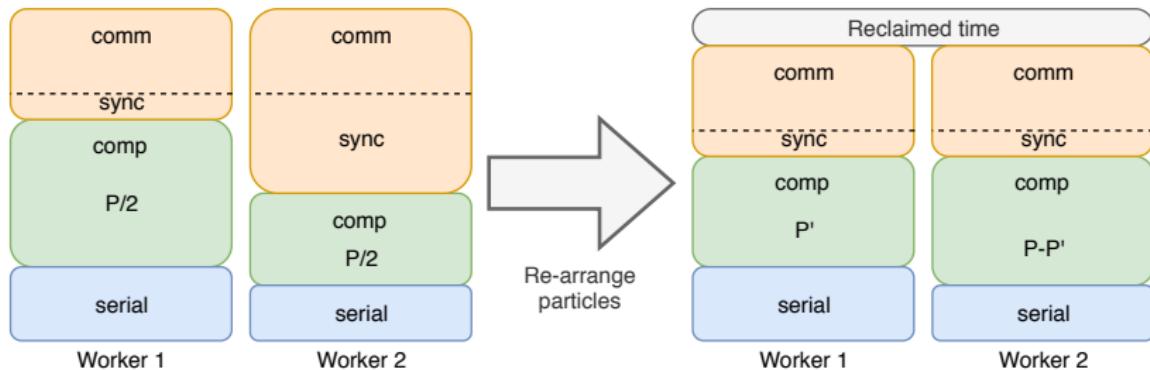
Testcase: PS



# Lost Time Due to Load Imbalance

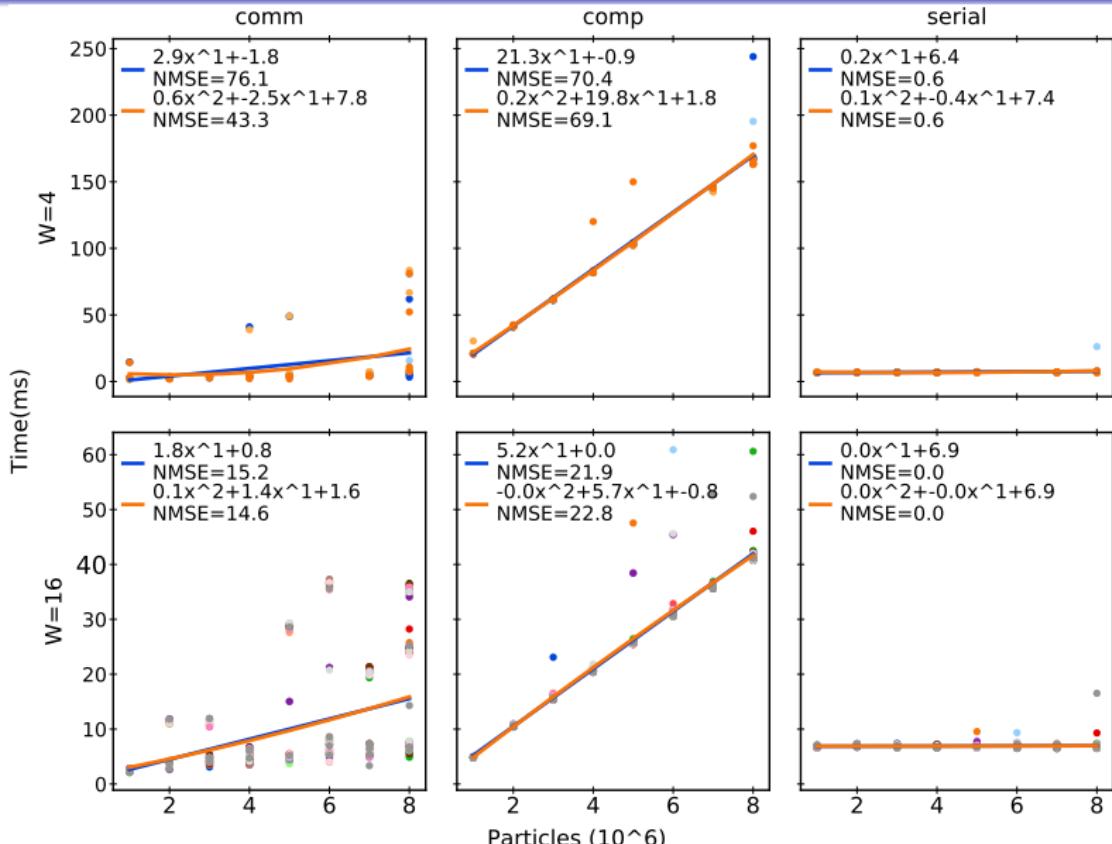


# Model Assumptions/ Conclusions

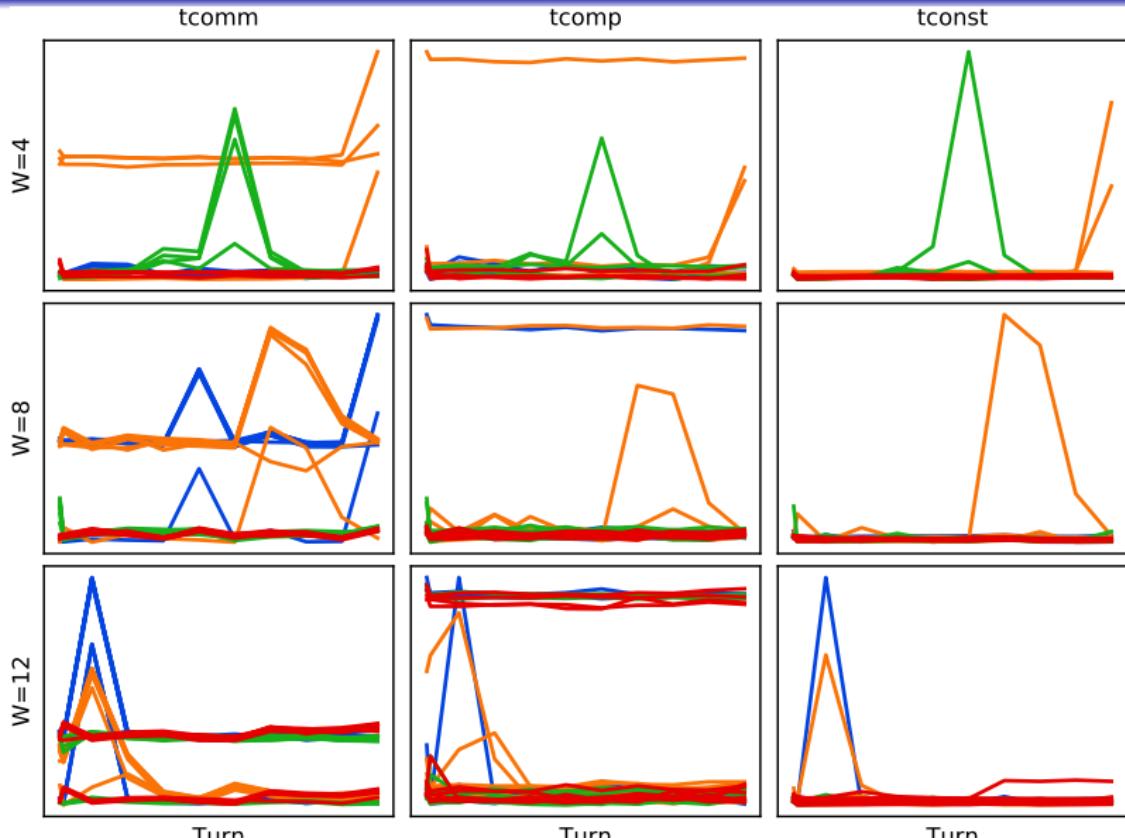


- ①  $T_{comm}^i(p) = \text{constant}$ ,  $T_{serial}^i(p) = \text{constant}$ .
- ②  $T_{comp}^i(p) = a^i \times p^i + b^i$ , where  $p$ : #particles,  $i$ : worker id.
- ③ Perfect LB  $\Leftrightarrow T_{serial}^i + T_{comp}^i = t \Leftrightarrow T_{sync}^i \rightarrow \min$ .
- ④ A worker exhibits the same behavior for long periods.
- ⑤  $T_{serial}^i$ ,  $T_{comp}^i$ ,  $a^i$ ,  $b^i$  can be calculated.

# Verifying the Assumptions 1,2 (PS)



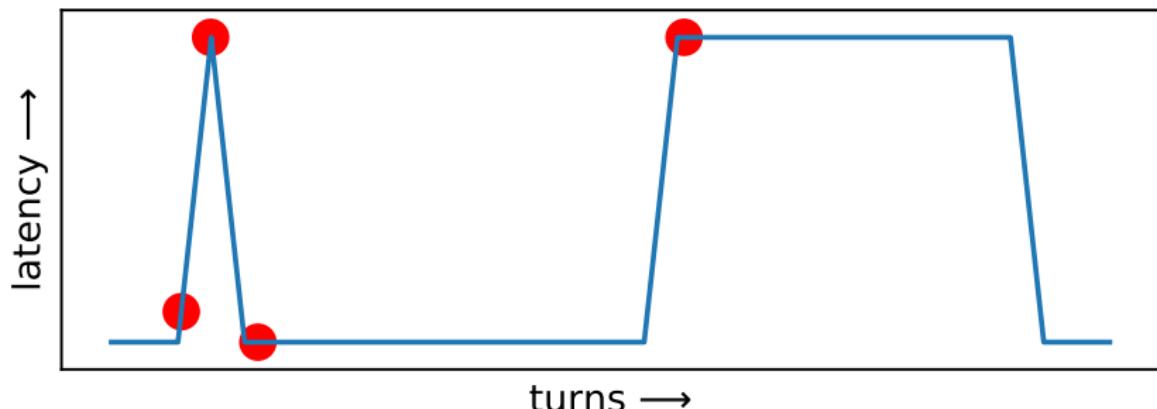
# Verifying the Assumptions 4 (PS)



# DLB Optimizations

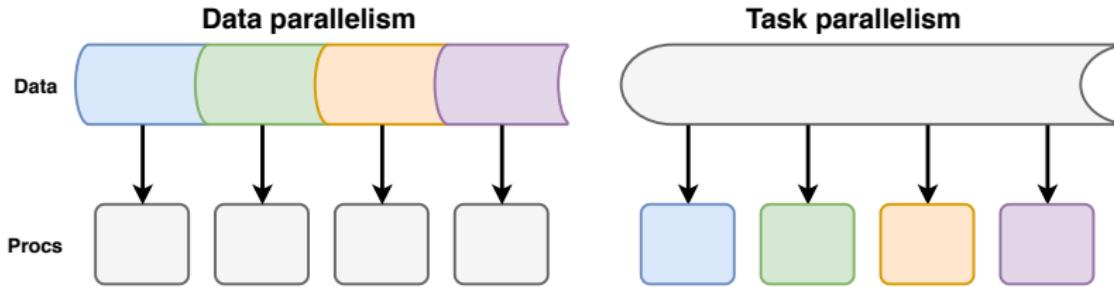
## Algorithmic optimizations

- Prioritize intra-node transactions.
- Minimize transactions: Sort by size, resolve the largest, repeat.
- Cut-off transaction size.
- Dynamically adjust the redistribution interval.
- Workers vote to initiate a redistribution.

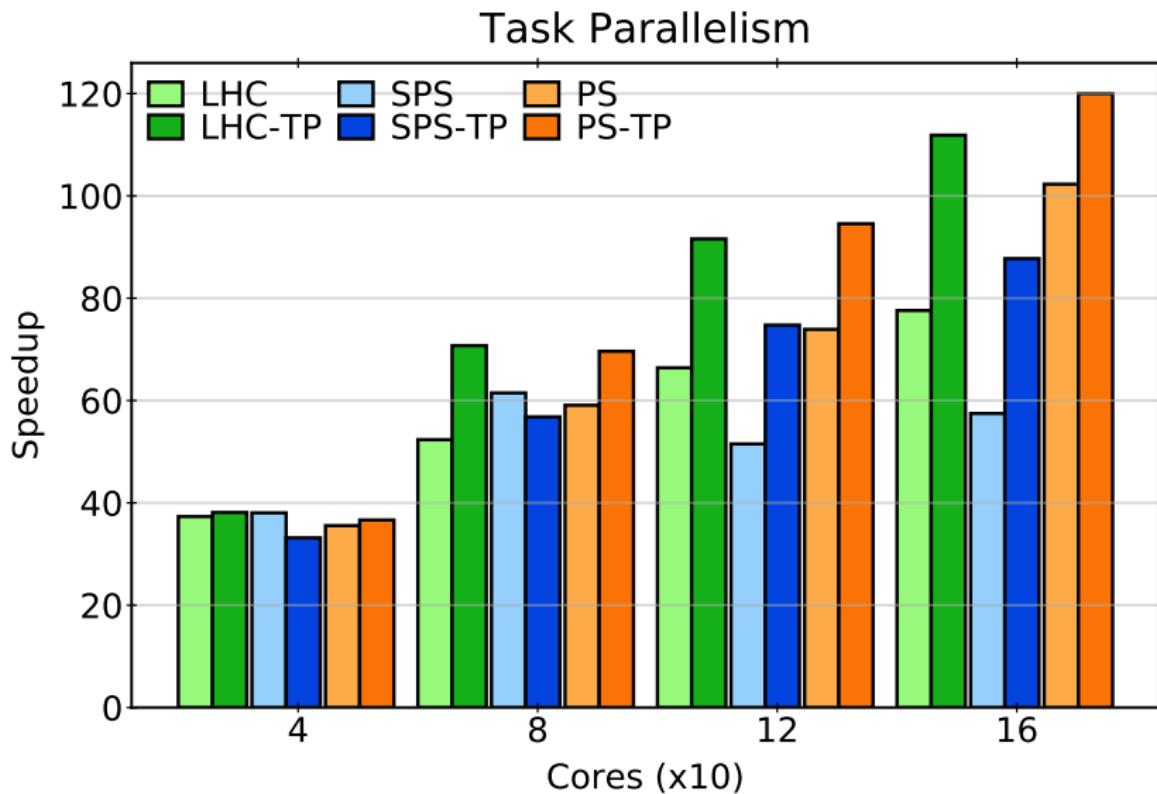


# Task Parallelism Opportunities

- Data Parallelism: Single operation, multiple data.
- Task Parallelism: Multiple operations, same data.
- The calculations of the induced voltage, the RF voltage and the beam feedbacks can be run in parallel.
- Neighboring workers: One calculates induced voltage, the other the RF voltage and the feedbacks.
- Tasks not equivalent → load balance.



# Task Parallelism, Experimental Results

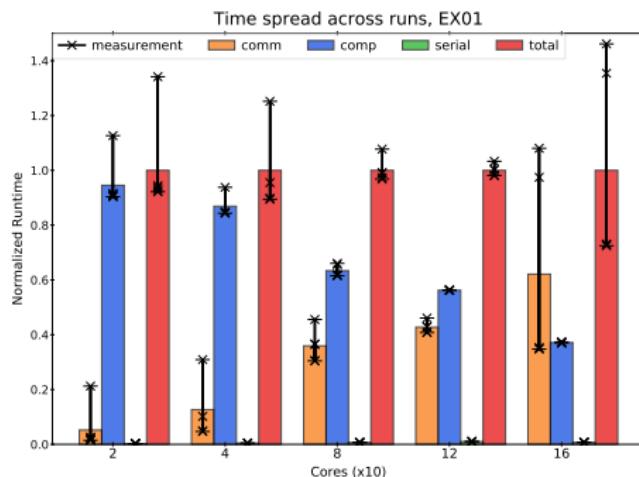


# Questions

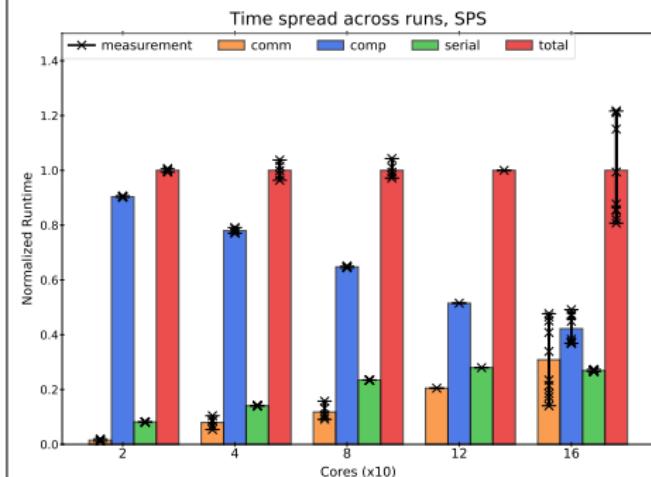


# Spread across runs(II)

Testcase: EX01

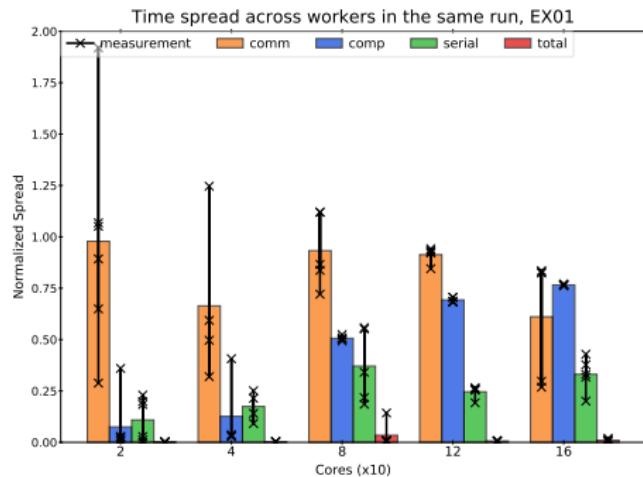


Testcase: SPS

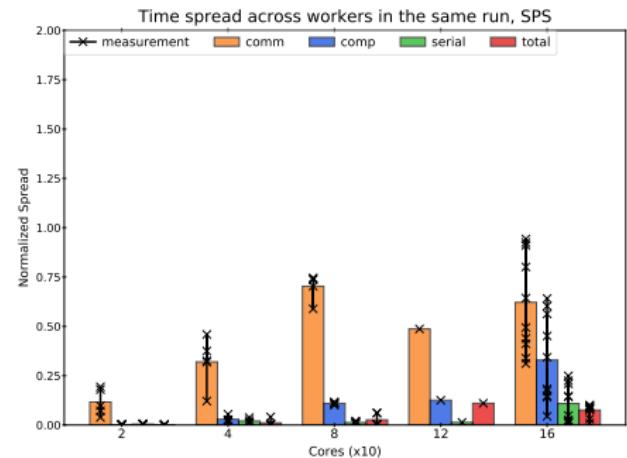


# Spread across workers(II)

Testcase: EX01



Testcase: SPS



# Verifying the Assumptions (LHC)

