RepEx TUU REMD performance optimization

Simulation setup

- All experiments performed on Stampede
- MD run in each dimension performs 6000 time-steps, which takes on average 136.64 seconds on one core, resulting in total MD simulation time on compute per cycle equal to 409.92 seconds
- All timings are averages of four simulation cycles
- For all simulation runs we always have enough cores to launch replicas concurrently (nr. cores >= nr. replicas)
- Number of replicas in each dimension is the same, resulting in total number of replicas being a third power of some number.

Simulation setup cont.

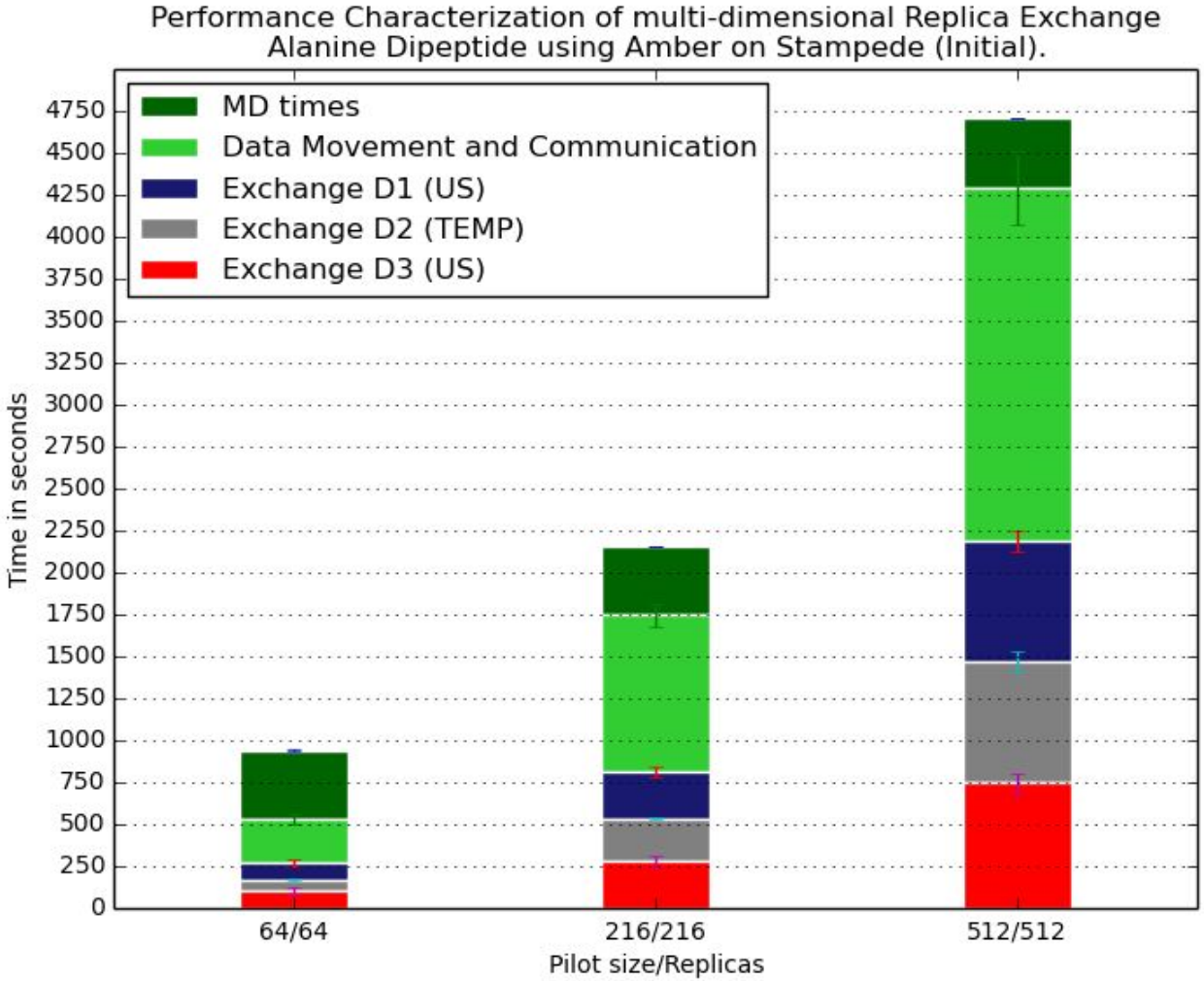
- We have used Alanine Dipeptide model consisting of 2881 atoms
- Order of dimensions is fixed: US → TEMP → US
- RADICAL-PILOT version - 0.33
- We have used Execution Pattern A

Initial implementation

- AMBER input files (.mdin; size: 334 bytes) are generated locally for each replica and then transferred to Stampede (this is done for each MD run)
- AMBER restraint files (.RST; size: 213 bytes) are generated locally and transferred to Stampede once (beginning of first cycle)
- Output files of exchange tasks - *matrix_column_d_c.dat* (size variable: 64 replicas - 331 bytes, 216 replicas - 962 bytes; 512 replicas - 2.2 Kilobytes)transferred from Stampede to laptop for final processing

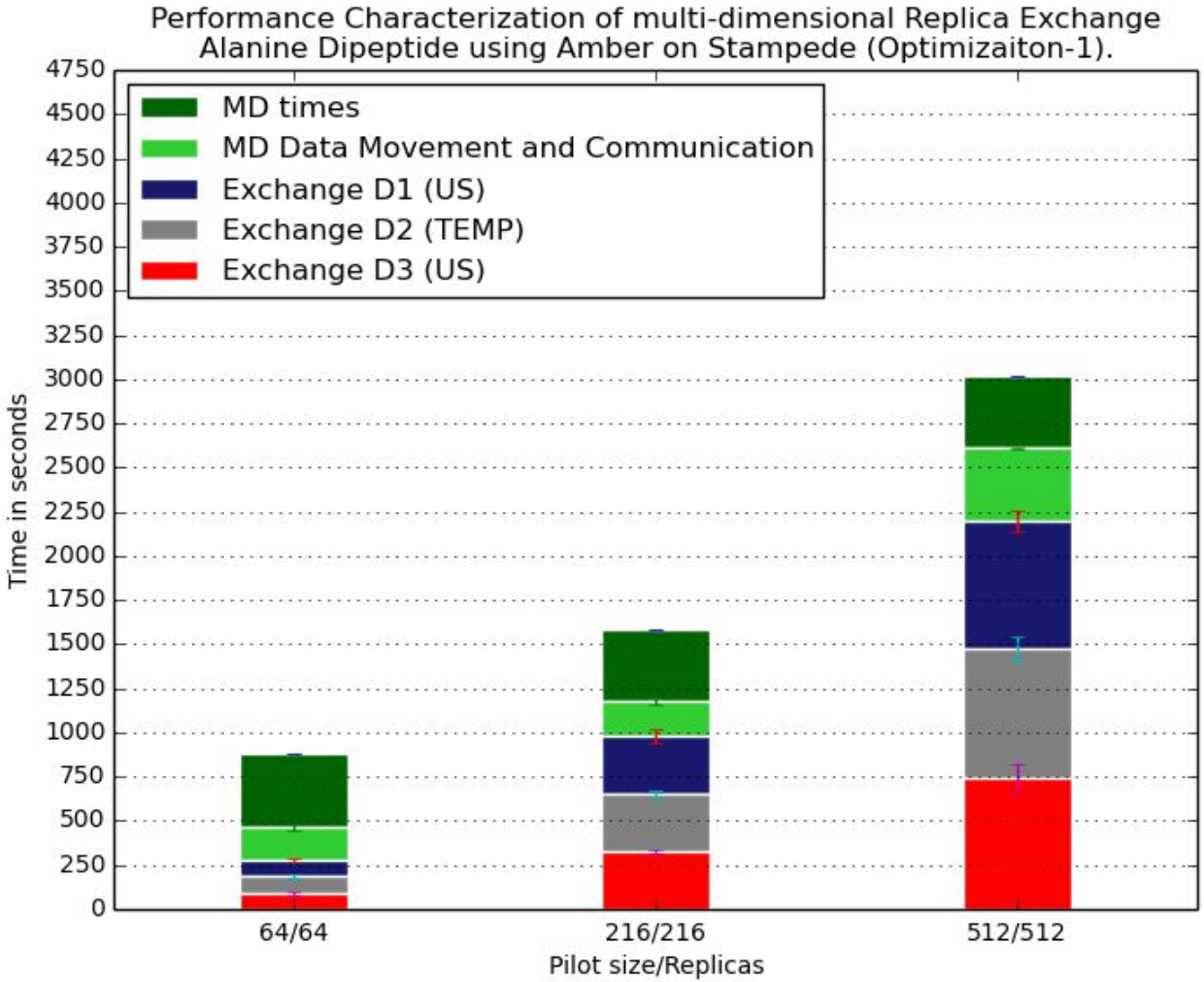
Initial implementation cont.

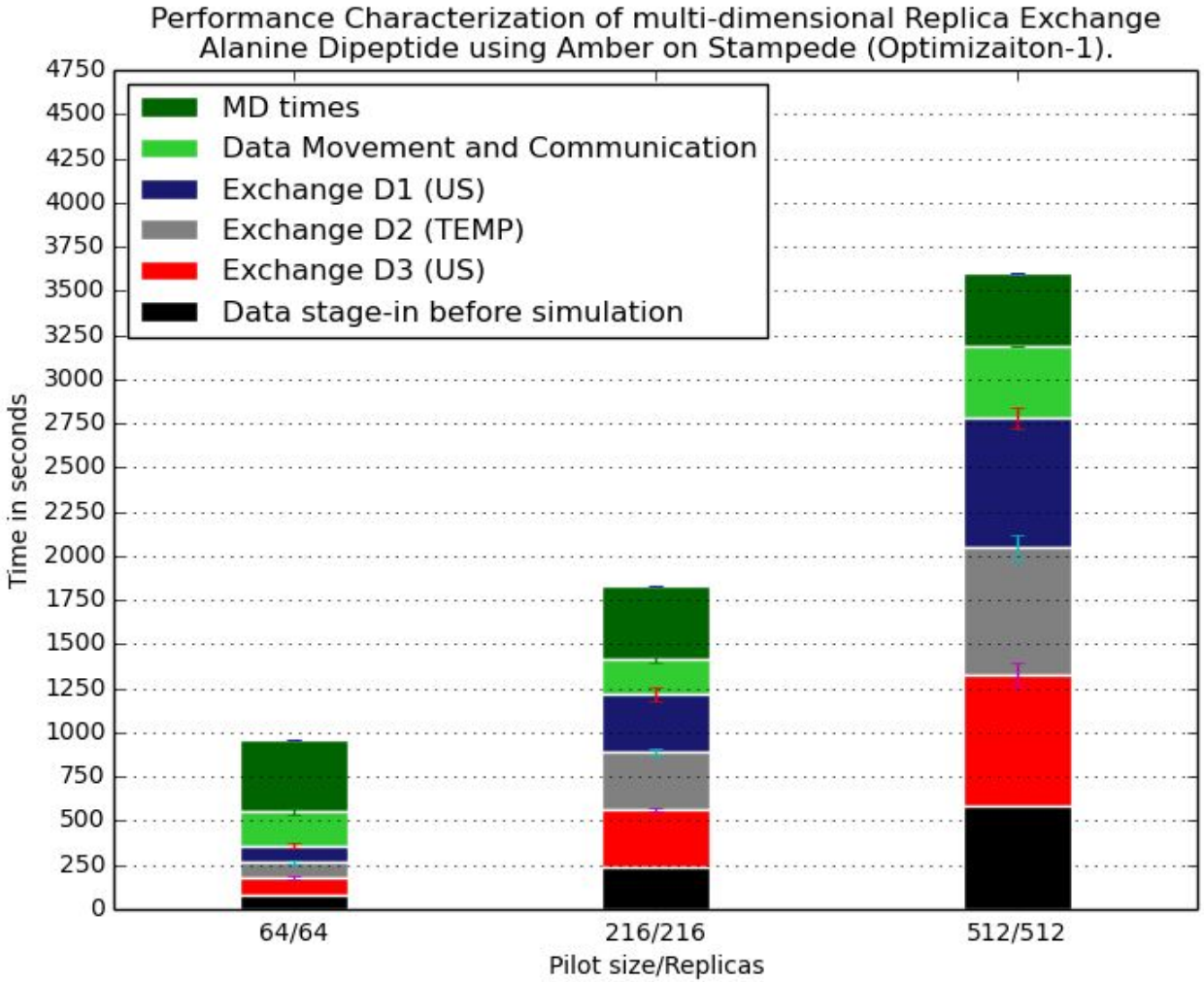
- Pairs of replicas for parameter exchanges are determined locally, after MD tasks and Exchange tasks for current run are done


Performance Characterization of multi-dimensional Replica Exchange
Alanine Dipeptide using Amber on Stampede (Initial).
MD times
Data Movement and Communication
Exchange D1 (US)
Exchange D2 (TEMP)
Exchange D3 (US)
Time in seconds
0
250
500
750
1000
1250
1500
1750
2000
2250
2500
2750
3000
3250
3500
3750
4000
4250
4500
4750
64/64
216/216
512/512
Pilot size/Replicas

Optimization-1

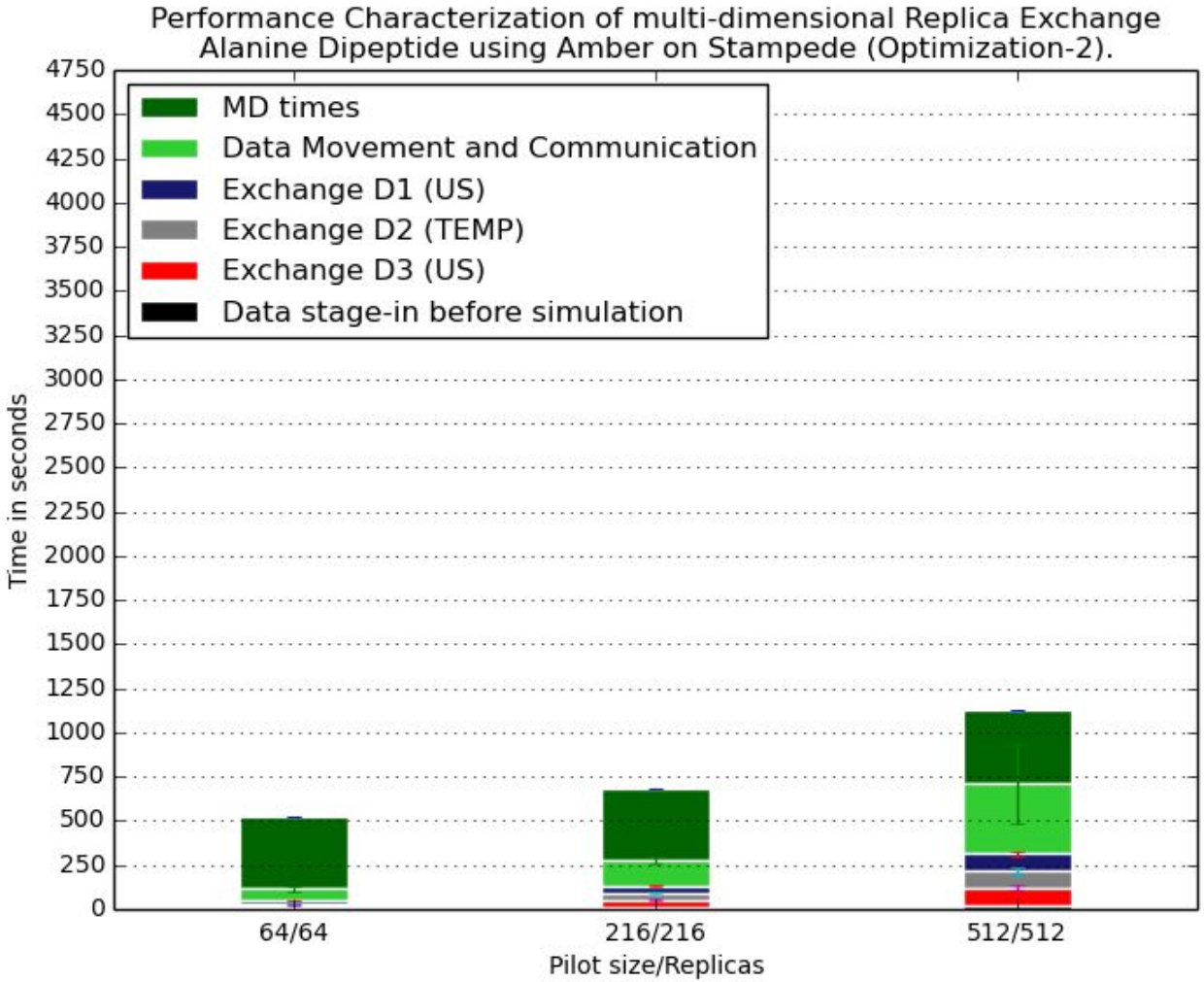
- Transfer to Stampede **only** input file template (shared) and script which generates input file, based on provided parameters (shared).
- Generate all input files remotely on Stampede during *pre_exec* step of MD tasks


Performance Characterization of multi-dimensional Replica Exchange
Alanine Dipeptide using Amber on Stampede (Optimizaiton-1).
MD times
MD Data Movement and Communication
Exchange D1 (US)
Exchange D2 (TEMP)
Exchange D3 (US)
Time in seconds
4750
4500
4250
4000
3750
3500
3250
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
64/64
216/216
512/512
Pilot size/Replicas


Performance Characterization of multi-dimensional Replica Exchange
Alanine Dipeptide using Amber on Stampede (Optimizaiton-1).
MD times
Data Movement and Communication
Exchange D1 (US)
Exchange D2 (TEMP)
Exchange D3 (US)
Data stage-in before simulation
Time in seconds
4750
4500
4250
4000
3750
3500
3250
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
64/64
216/216
512/512
Pilot size/Replicas

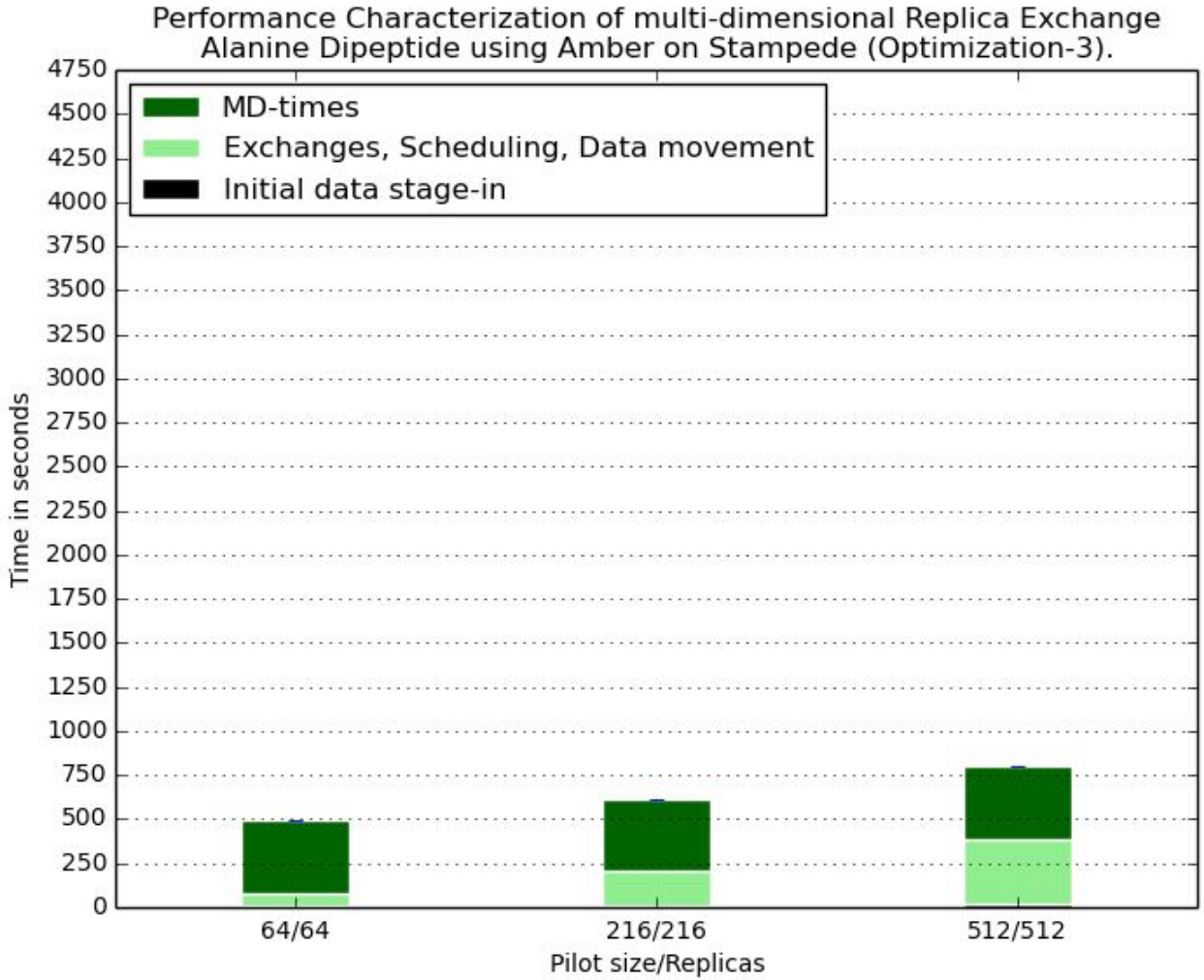
Optimization-2

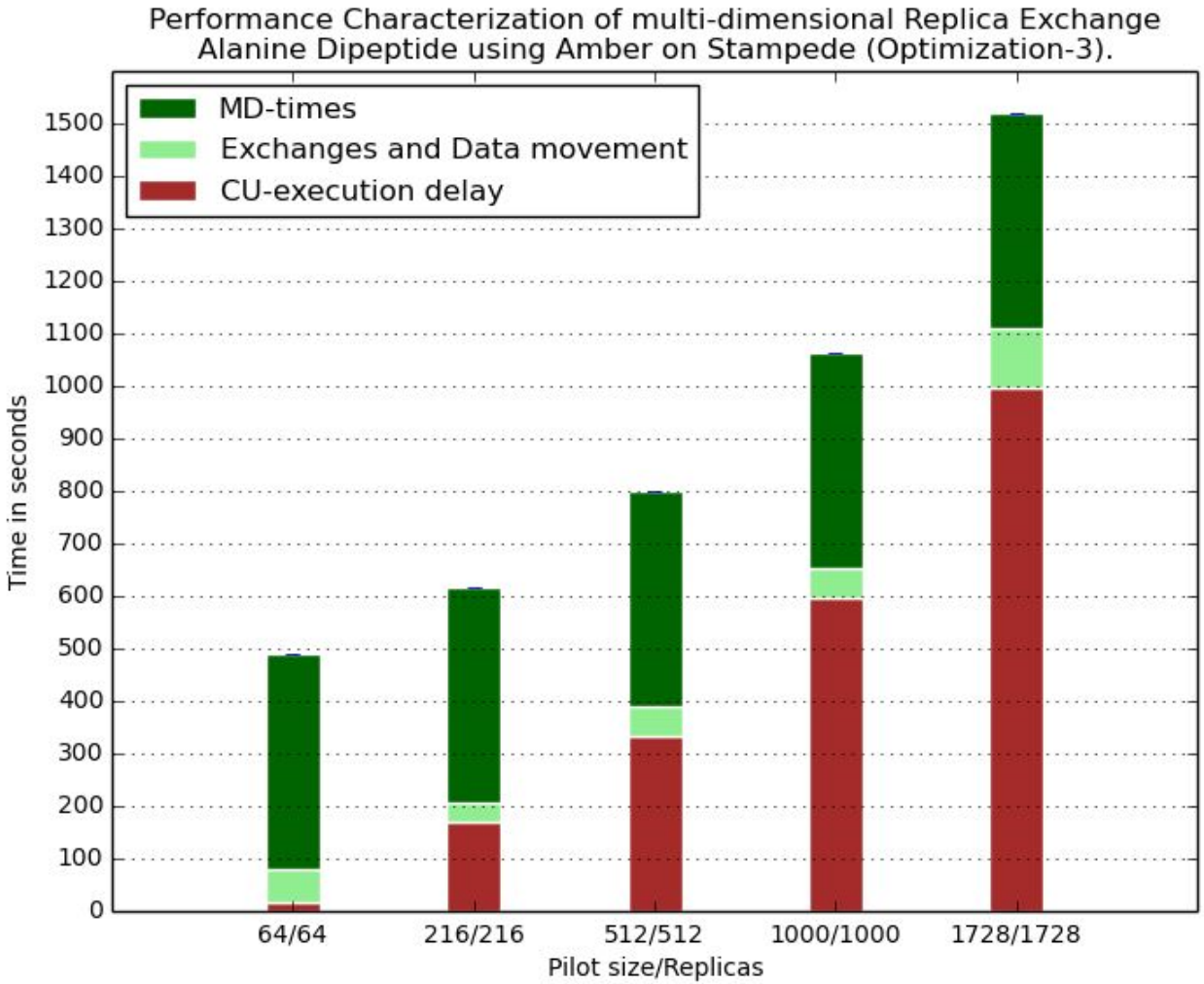
- Transfer to Stampede **only** restraint file template (shared) and script which generates restraint file, based on provided parameters (shared).
- Generate all restraint files remotely on Stampede during *pre_exec* step of MD tasks (during 1 cycle only)
- Determine pairs of replicas for exchange remotely and transfer back to laptop only **one** file with replica ids, instead of **N** *matrix_column_d_c.dat* files. This requires to run remotely part of application which processes column files and using Gibbs sampling determines pairs of replicas for exchange


Performance Characterization of multi-dimensional Replica Exchange
Alanine Dipeptide using Amber on Stampede (Optimization-2).
MD times
Data Movement and Communication
Exchange D1 (US)
Exchange D2 (TEMP)
Exchange D3 (US)
Data stage-in before simulation
Time in seconds
4750
4500
4250
4000
3750
3500
3250
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
64/64
216/216
512/512
Pilot size/Replicas

Optimization-3

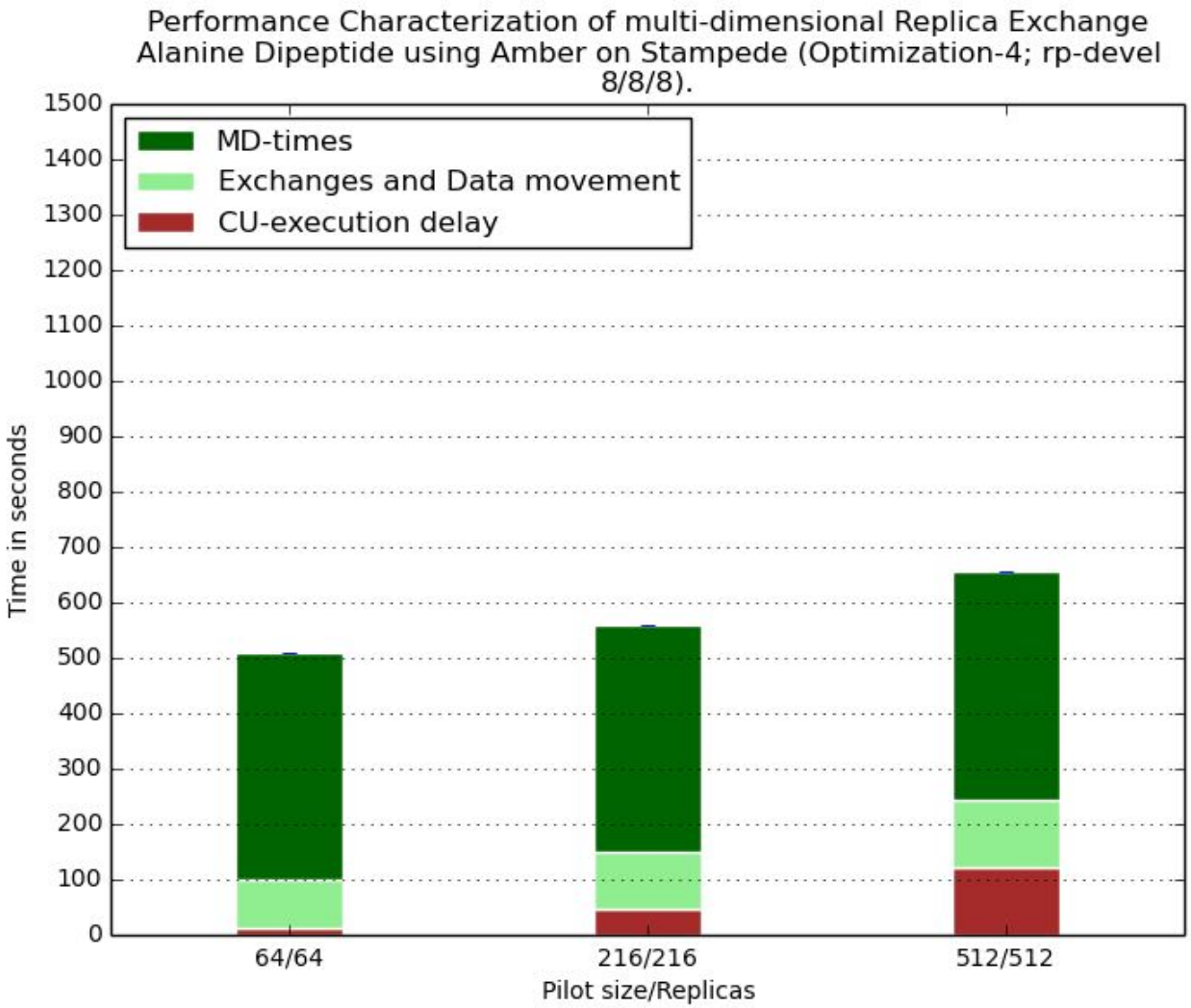
- Instead of creating a separate set of CUs for Exchange tasks, run Exchange tasks as *post_proc* step of MD tasks


Performance Characterization of multi-dimensional Replica Exchange
Alanine Dipeptide using Amber on Stampede (Optimization-3).
MD-times
Exchanges, Scheduling, Data movement
Initial data stage-in
Time in seconds
4750
4500
4250
4000
3750
3500
3250
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
64/64
216/216
512/512
Pilot size/Replicas


Performance Characterization of multi-dimensional Replica Exchange
Alanine Dipeptide using Amber on Stampede (Optimization-3).
MD-times
Exchanges and Data movement
CU-execution delay
Time in seconds
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
64/64
216/216
512/512
1000/1000
1728/1728
Pilot size/Replicas

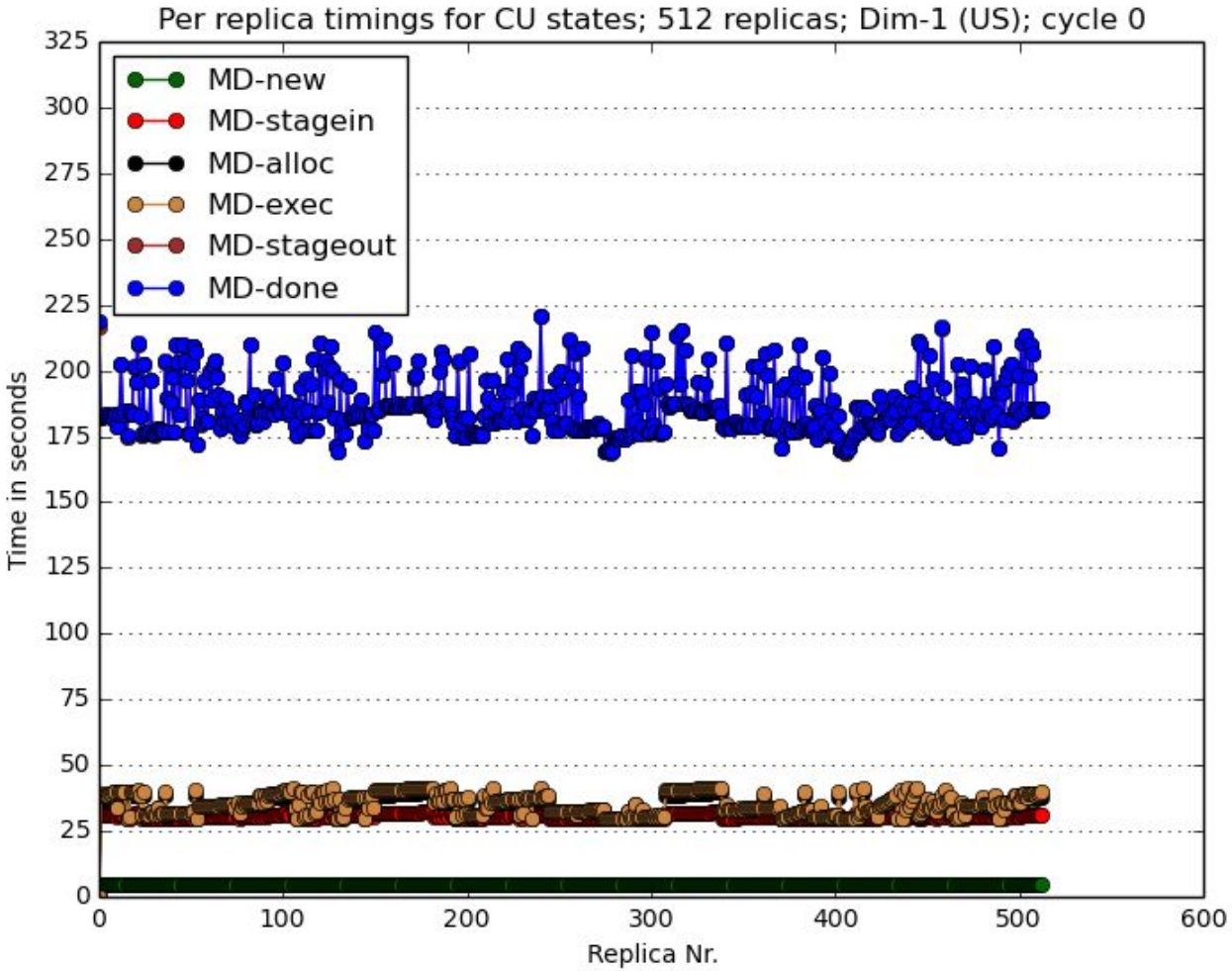
Optimization-4

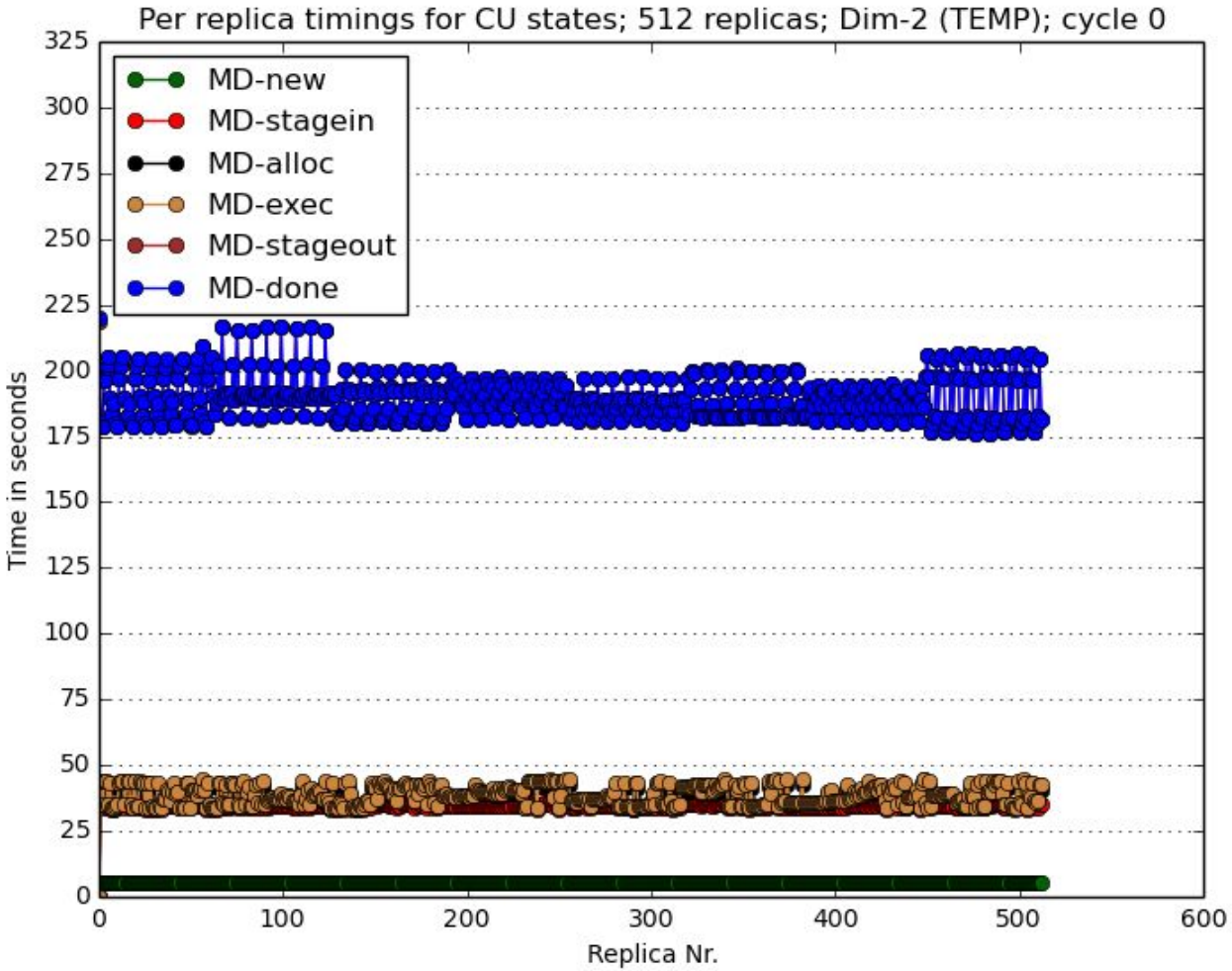
- Using RP-devel branch (v0.33-143-g1529c5d@devel)
- Bulk submissions
- StageinWorkers = 8; ExecWorkers = 8; StageoutWorker = 8

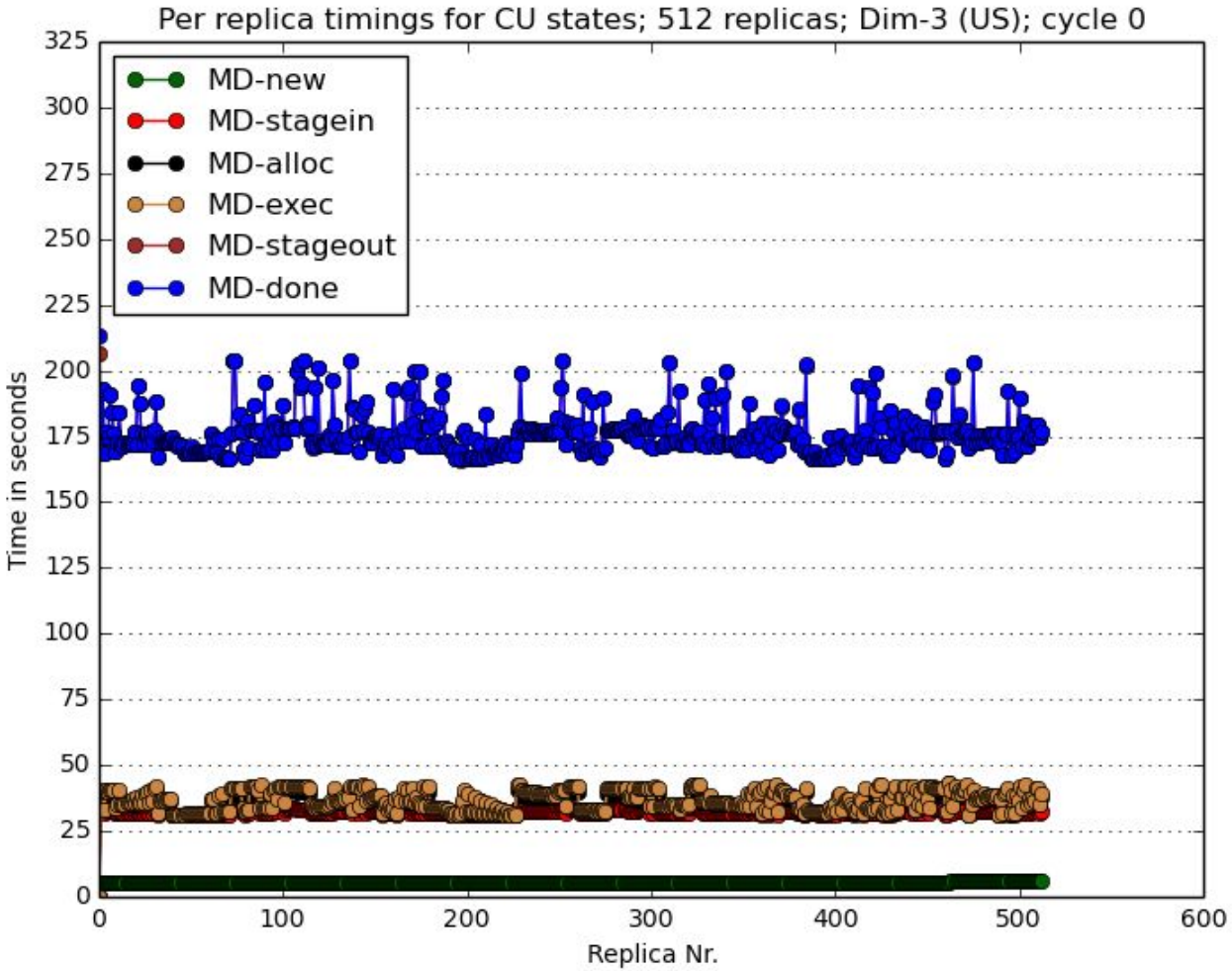

Performance Characterization of multi-dimensional Replica Exchange Alanine Dipeptide using Amber on Stampede (Optimization-4; rp-devel 8/8/8).
MD-times
Exchanges and Data movement
CU-execution delay
Time in seconds
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
64/64
216/216
512/512
Pilot size/Replicas

Closer look

- For Opt-3 and Opt-4 we run “global exchange script” concurrently with individual exchange tasks and this script is launched **before** these tasks. This allows to overlap execution of “global exchange script” and individual exchange tasks, but requires one extra core. In principle we can run “global exchange script” **after** all individual exchange tasks are finished without this extra core.


Per replica timings for CU states; 512 replicas; Dim-1 (US); cycle 0
MD-new
MD-stagein
MD-alloc
MD-exec
MD-stageout
MD-done
Time in seconds
Replica Nr.
0
25
50
75
100
125
150
175
200
225
250
275
300
325
0
100
200
300
400
500
600


Per replica timings for CU states; 512 replicas; Dim-2 (TEMP); cycle 0
MD-new
MD-stagein
MD-alloc
MD-exec
MD-stageout
MD-done
Time in seconds
Replica Nr.
0
25
50
75
100
125
150
175
200
225
250
275
300
325
0
100
200
300
400
500
600


Per replica timings for CU states; 512 replicas; Dim-3 (US); cycle 0
MD-new
MD-stagein
MD-alloc
MD-exec
MD-stageout
MD-done
Time in seconds
Replica Nr.
0
25
50
75
100
125
150
175
200
225
250
275
300
325
0
100
200
300
400
500
600

Some observations (512 replicas)

- Overhead decomposition:
  - Total overhead: 245.02
  - CU startup delay: 123.05
  - Exchanges and data movement: 121.97