

Co-scheduling Ensembles of In Situ Workflows

Tu Mai Anh Do*, Loïc Pottier*, Rafael Ferreira da Silva[‡], Frédéric Suter[‡], Silvina Caíno-Lores[§],
Michela Taufer[§], Ewa Deelman*

*Information Sciences Institute, University of Southern California, Marina Del Rey, CA, USA

[‡]Oak Ridge National Laboratory, Oak Ridge, TN, USA

[§]University of Tennessee at Knoxville, Knoxville, TN, USA

{tudo,lpottier,deelman}@isi.edu, {silvarf,suterf}@ornl.gov, {mtaufer,scainolo}@utk.edu

Abstract—Molecular dynamics (MD) simulations are widely used to study large-scale molecular systems. HPC systems are ideal platforms to run these studies, however, reaching the necessary simulation timescale to detect rare processes is challenging, even with modern supercomputers. To overcome the timescale limitation, the simulation of a long MD trajectory is replaced by multiple short-range simulations that are executed simultaneously in an ensemble of simulations. Analyses are usually co-scheduled with these simulations to efficiently process large volumes of data generated by the simulations at runtime, thanks to in situ techniques. Executing a workflow ensemble of simulations and their in situ analyses requires efficient co-scheduling strategies and sophisticated management of computational resources so that they are not slowing down each other. In this paper, we propose an efficient method to co-schedule simulations and in situ analyses such that the makespan of the workflow ensemble is minimized. We present a novel approach to allocate resources for a workflow ensemble under resource constraints by using a theoretical framework modeling the workflow ensemble’s execution. We evaluate the proposed approach using an accurate simulator based on the WRENCH simulation framework on various workflow ensemble configurations. Results demonstrate the significance of co-scheduling simulations and in situ analyses that couple data together to benefit from data locality, in which inefficient scheduling decisions can lead up to a factor 30 slowdown in makespan.

Index Terms—workflow ensemble, in situ, co-scheduling, molecular dynamics, high-performance computing

I. INTRODUCTION

Molecular dynamics (MD) is one of the scientific simulations that benefit most from ever-increasing computational capabilities offered by modern high-performance computing (HPC) platforms. MD simulations serve as a productive method to observe relevant processes of a molecular system at atomic resolution. MD simulations discover the motion of the system through computing atomic positions that evolve over time. MD simulations regularly generate large amount of data that needs to be stored efficiently to disks. However, I/O capabilities have not evolved as fast as compute capabilities in contemporary HPC infrastructures, which causes an important performance bottleneck at scale [1] due to large amount of data generated by the simulations that cannot be stored efficiently to disks. Consequently, the community has developed different approaches to overcome the I/O bottleneck—in this work, we focus on a particular one called in situ. The in situ paradigm shifts the data processing phase from traditional post-processing to iterative processing following a producer-

consumer approach. Basically, instead of storing data produced by simulations on disks, simulations send data periodically using faster memories (e.g., DRAM, SSD) to in situ analysis jobs that run concurrently to the simulations. Performing analyses in situ does not only help to obtain on-the-fly insights into the molecular system at runtime, but it also reduces the burden on the file system.

Interesting molecular events require observing at a long enough simulation timescale, which is compute-intensive for simulating large-scale molecular systems. To overcome this timescale problem, a family of enhanced-sampling methods utilizes multiple short-range simulations running within *ensembles* to efficiently explore different regions of the conformational space. Ensemble-based computational methods [2]–[5] are gaining popularity in many scientific domains using computational simulations, where computations are executed as jobs into a *workflow ensemble*. A workflow ensemble [6] is usually composed of a number of inter-related workflows, for example identical workflows of simulation-analysis pipeline with different input parameters exploring the solutions space. We consider the execution of ensembles of in situ workflows in which concurrent jobs, i.e. simulations and analyses, are tightly coupled using in situ processing, thereby usually co-scheduled together to efficiently process data at runtime. However, sustaining complex co-scheduling for massive workflow ensembles of concurrent applications will require sophisticated orchestration tailored for the management of workflow ensembles [7], [8].

In this paper, we aim to guide the above emerging orchestration runtime. The key challenge when orchestrating the workflow ensembles at scale is to efficiently schedule the simulations and in situ analyses as they continuously exchange data. The intersection of workflow ensembles and in situ is particularly challenging as in situ itself has intricate, complex communication patterns, which is further convoluted by the number of concurrent executions at scale. Intuitively, co-scheduling the analyses with their associated simulation on the same computing resources improves data locality, but required resources may not be consistently available. Thus, upon resource constraint, how do we co-schedule workflow ensembles on a HPC machine? A key distinguishing factor is to efficiently generate large ensembles to optimize scientific discoveries. Another major challenge is to determine resource requirements such that the resulting performance of executing

them together in the workflow ensemble is maximized. Therefore, we design efficient co-scheduling strategies and resource assignments for the workflow ensemble of simulations and in situ analyses. Our contributions are as follows:

- We develop a mathematical model that characterizes iterative execution and data coupling behavior of simulations and in situ analyses in a workflow ensemble.
- We use this model to derive optimal co-scheduling strategies and compute resource allocations for the simulations and in situ analyses in the workflow ensemble under constraints of the available computing resources.
- We evaluate our proposed scheduling solutions using an accurate simulator by studying different co-scheduling scenarios on various configurations of the workflow ensembles.

II. RELATED WORK

Designing scheduling algorithms for in situ workflows such that available resources are efficiently utilized is challenging as optimizing individual workflow components does not ensure that the end-to-end performance of the workflow is optimal. The larger the number of components the workflow has, the larger the multi-parametric space needed to explore. Specifically, workflow ensembles consist of many simulations and in situ analyses running concurrently, thus the combination of their data coupling behaviors significantly increases the complexity of the scheduling decisions. Few efforts have been attempted to solve this problem on a single in situ workflow. Malakar *et al.* formulated the in situ coupling between the simulation and the analysis as a mixed-integer linear programming problem [9] and derived the frequency to schedule the analysis executed in situ on different set of nodes. Aupy *et al.* proposed a greedy algorithm [10] to schedule a set of in situ analyses on available resources with memory constraints. Taufer *et al.* introduced a 2-step model [11] to predict the frequency of frames to be analyzed in situ such that computing resources are not underutilized while the simulation is not slowed down by the analysis. Our proposed approach for scheduling is not restricted to a single in situ coupling.

Co-scheduling strategies have been recently proposed to deliver better resource utilization and increasing overall application throughput [12]–[14]. However, only few works incorporated co-scheduling into the in situ workflows, where components (simulations and analyses) in the workflow are tightly coupled together. Sewell *et al.* proposed a thorough performance study [15] for co-scheduling in situ components that leverages memory system for data staging with low latencies. Aupy *et al.* introduced an optimization-based approach [16] to determine the optimal fraction of caches and cores allocated to in situ tasks. Due to the lack of capability to submit a batch of concurrent applications with conventional scheduler, co-scheduling was not fully-supported for the workflow ensembles at production scale. To overcome this limitation, Flux [7] was designed as a hierarchical resource manager to meet the needs of co-scheduling large-scale ensemble of various simulations and in situ analyses. To the best of our

knowledge, this paper is the first effort that addresses co-scheduling problem for in situ workflows at ensemble level.

III. MODEL

The goal of this paper is to study the scheduling of complex ensembles of in situ workflows [17] on parallel machines, where every workflow within an ensemble has several concurrent jobs coupled together using in situ processing, either be a *simulation* or an *analysis*. Unfortunately, one might not have enough compute nodes to run concurrently all simulations and analyses and, will have to make scheduling choices: (i) which simulation and analysis should run together on the same resources and (ii) how much resources should be allocated to each.

Our approach. The core of this work is to propose a powerful but versatile model to express complex data dependencies between simulations and in situ analyses in a workflow ensemble, and based on that design appropriate scheduling strategies. The chosen approach is to model the scheduling problem under simplified conditions where we make several assumptions to reduce the large space of exploration in an ensemble of in situ workflows. Then, we demonstrate that solutions designed for the simplified world, also perform well in using realistic settings.

A. Couplings

A simulation S_i represents a job in an ensemble that simulates a state of interest in the molecular system. An analysis, denoted as A_j , represents a computational kernel that is suitable to analyze in situ data produced by a simulation, e.g. transforming high-dimensional molecular structures to eigenvalue metadata [18]. Regularly, the output resulted by in situ analyses are later post-processed, e.g. synthesized to infer conformational changes of the molecular system in a post-hoc analysis. However, in the scope of this paper, we focus on the part of the simulation-analysis pipeline where they are executed simultaneously in an ensemble as it is more challenging to manage. We denote the set of simulations and analyses in an ensemble as S and A respectively. In this work, each workflow of an ensemble is composed of one simulation and at least one analysis. The analyses are executed in parallel with the simulation and periodically processed simulation data generated after each simulation steps in an iterative fashion. This essence of in situ processing expresses the important notion of *coupling*.

Coupling indicates which simulation and analyses have a data dependency, i.e. data produced by the simulation is analyzed by the corresponding analyses. Scientists usually decides beforehand which analyses have to process the data from a given simulation, we call this mapping a *coupling* (see Figure 1). Therefore, couplings are decided beforehand as an input of the problem. More formally, a coupling $p : S \rightarrow A$ defines which analyses in A are coupled with which simulation in S , then for every $S_i \in S$, let $p(S_i)$ be the set of analyses that couple data with the simulation S_i . For example, in Figure 1, we have $p(S_1) = \{A_1, A_3\}$. Now that we have defined the

notion of couplings, we need to define different notions around *co-scheduling*.

B. Co-scheduling

The notion of co-scheduling is at the core of this work, *co-scheduling* is defined as running multiple applications concurrently on the same set of compute nodes, each application using a fraction of the number of cores per node. We assume the total number of cores of each application is distributed evenly among the compute nodes the application is co-scheduled on, which is similar to the manner of distributing resources in emerging schedulers, e.g. Flux [7], for co-scheduling concurrent jobs of a massive ensemble in next-generation computers. In this paper, we distinguish two scenarios in which we map the simulations and analyses to computational resources:

- *Co-scheduling*: S_i and A_j run on the same compute nodes and thus have access to the shared memory;
- *In transit*: S_i and A_j run on two different set of nodes and sending data from S_i to A_j involves network communications.

In an attempt to avoid performance degradation in co-scheduling, in this work, we consider that simulations, which are compute-intensive, are not co-scheduled together [10]

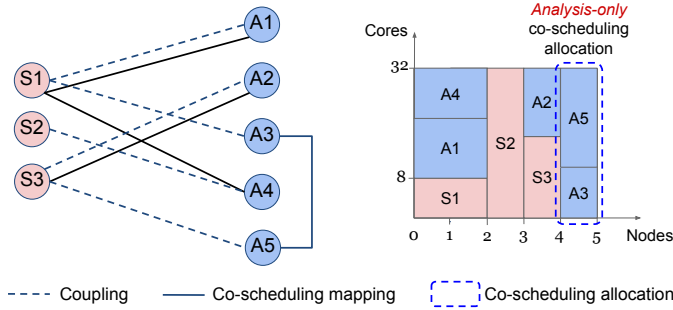


Fig. 1: Illustration of co-scheduling and coupling notions.

Co-scheduling mapping. A co-scheduling mapping m defines how S and A are co-scheduled together. For a given mapping m , we denote by $m(S_i)$ the set of analyses co-scheduled with S_i . For example, with the mapping illustrated in Figure 1, $m(S_1) = \{A_1, A_3\}$ and $m(S_3) = \{A_2\}$.

Assumption 1. In this model, we ignore co-scheduling interference, such as cache sharing, that could degrade the performance when multiple applications share resources [19].

Co-scheduling allocation. A co-scheduling allocation represents a set of applications which are co-scheduled together (i.e., share computing resources). Let $\langle \{S_i, A_j\} \rangle$ denote a co-scheduling allocation where simulation S_i is co-scheduled with analysis A_j . Recall that, we have at most one simulation per allocation, but we can have multiple analyses. Formally, for a given co-scheduling mapping m , a simulation S_i and analyses $m(S_i)$ that are co-scheduled with S_i , the corresponding

co-scheduling allocation is denoted by $\langle S_i \cup m(S_i) \rangle$. However, not all analyses are required to be co-scheduled with a simulation, there exists co-scheduling allocations in which only analyses are co-scheduled together, called *analysis-only co-scheduling allocations*. For example, in Figure 1, $\langle \{A_3, A_5\} \rangle$ is an analysis-only co-scheduling allocation. A co-scheduling allocation also determines the amount of computer resources assigned for each co-scheduled application. Let n_x denote the number of nodes assigned to application x . For instance, if S_i and A_j are co-scheduled on the same co-scheduling allocation, then $n_{S_i} = n_{A_j}$. Let c_x be the number of cores per node assigned to x . As each co-scheduled application takes a portion of cores in the total of C cores a node has, then for every co-scheduling allocation $\langle X \rangle$, we have $\sum_{x \in X} c_x \leq C$. In Figure 1, since A_1 and A_4 are co-scheduled with S_1 on a co-scheduling allocation spreading in 2 compute nodes, then $n_{S_1} = n_{A_1} = n_{A_4} = 2$. And since $c_{S_1} = 8$, then S_1 occupies 8 cores each node, so 16 cores in total.

C. Execution platform

We consider a traditional parallel computing platform with N identical compute nodes, each node has C identical cores, a shared memory of size M and a bandwidth of B per node. In addition, we consider that the interconnect network between nodes is a fully-connected topology, hence the communication time is the same for any pairs of nodes.

D. Application model

Because of the in situ execution, simulations and analyses are executed iteratively, in which iterations exhibit consistent behavior [20]. Specifically, the simulation executes certain compute-intensive tasks each iteration and then generates output data. The analysis, in every iteration, subscribes the data produced by the simulation and performs several specific computations. Since the complexity of computation and the amount of data are identical across iterations, the time to execute one iteration is approximately the same across iterations. Based on this reasoning and, by assuming the simulations and analyses have the same number of iterations (n_{steps}), we simply consider the execution time of a single iteration as the entire execution time can be derived from multiplying the time taken by a single iteration by n_{steps} [20].

Computation model. Let us now denote by $t(x)$ the time to execute one iteration of the iterative application x (x can either be a simulation or an analysis). For readability, we pay particular attention to the class of perfectly parallel applications, even though the approach can be generalized to other models, such as Amdahl's Law [21].

Assumption 2. Every simulation and analysis follow a perfectly parallel speedup model [22].

In other words, a job x that runs in time $t_x(1)$ on one core will run in time $t_x(1)/c$ when using c cores. Thus, we can express the time to execute one iteration of the simulation as follows

$$t(S_i) = \frac{t_{S_i}(1)}{n_{S_i} c_{S_i}}, \quad (1)$$

where S_i is running on n_{S_i} compute nodes with c_{S_i} cores each node.

Communication model. In in situ processing, data is usually stored on node-local storage [23] or local memory [24] to reduce the overhead associated with data movement. In this work, we process data residing in memory to allow near real-time data analysis. Data communications are modeled as follows: (i) S_i writes its results into local memory of the nodes it runs on, and (ii) each analysis coupled with S_i reads from that memory the data it needs whether locally or remotely depends on A_j is co-scheduled with S_i or not.

Assumption 3. *Intra-node communications (i.e., shared memory) are considered negligible compared to inter-node communications (i.e., distributed memory).*

In this model, we consider that writing/reading to/from the local memory of a node (e.g., the RAM) is negligible, therefore the cost of communicating data between simulation and analysis co-scheduled together is approximately close to zero. However, communications going over the interconnect have to be modeled. For example, in Figure 1, A_3 , A_4 and A_5 are not co-scheduled with the simulations they couple with, hence they incur overhead of remote transfers. To model these inter-node communications, we adopt a classic linear bandwidth model [25] where it takes $\frac{S}{B}$ to communicate a message of size S , where B is the bandwidth. Let V_{A_j} be the amount of data received and processed by A_j . Remember that A_j runs on n_{A_j} compute nodes with c_{A_j} core per node. Then, based on the linear bandwidth model, the I/O time of A_j can be expressed as $V_{A_j}/B n_{A_j}$, where B is the maximum bandwidth of one compute node. Even though the bandwidth is actually variable as it is shared by multiple I/O operations during the execution, we use the maximum bandwidth of a compute node for an optimistic consideration.

Execution time. Finally, we combine computation and communication model to get actual time to execute one iteration:

$$t(A_j) = \begin{cases} \frac{t_{A_j}(1)}{n_{A_j} c_{A_j}} & \text{if } A_j \text{ is co-scheduled with} \\ & \text{the simulation it couples with;} \\ \frac{t_{A_j}(1)}{n_{A_j} c_{A_j}} + \frac{V_{A_j}}{B n_{A_j}} & \text{otherwise;} \end{cases} \quad (2)$$

while $t(S_i)$ is simply Equation (1) as communications to local memory are free (cf. Assumption 3). Table I summarizes the notation introduced in this section.

E. Makespan

We have defined the execution time for the simulation and the analysis and, armed with the notion of *co-scheduling mapping* previously discussed, we are ready to define the makespan of an ensemble of in situ workflows. Let n_{steps} denote the number of iterations, recall that we assume the simulations and analyses in an ensemble have the same number of iterations. The makespan can now be expressed as follows:

$$\text{MAKESPAN} = \max_{S_i \in S, A_j \in A} (t(S_i), t(A_j)) \times n_{steps} \quad (3)$$

Notation	Description
S_i, A_j	Set of simulations and analyses ($1 \leq i, j \leq n$)
N, C	Total number of nodes and number of cores per node
B	Maximum bandwidth per node
$p(S_i)$	Set of analyses that couple data with simulation S_i
V_{A_j}	Amount of data received by analysis A_j
$m(S_i)$	Set of analyses co-scheduled with simulation S_i
$\{S_i, A_j\}$	Co-scheduling allocation where S_i and A_j are co-scheduled
n_x, c_x	Number of nodes and cores/node assigned to x
$t_x(1)$	Time to execute one iteration of x on one core
$t(x)$	Actual time to execute one iteration of x on n_x, c_x
P^{NC}	Set of analyses co-scheduled on analysis-only allocations

TABLE I: Notations.

We are now ready to define different scheduling problems tackled in this work such that this makespan is minimized.

IV. SCHEDULING PROBLEMS

We consider that simulations and analysis are moldable jobs [26] (i.e., resources allocated to a job are determined before its execution starts). In addition, all the results presented in this section are using rational amount of resources, i.e. number of nodes and cores are rational. Our approach is to design optimal, but rational solutions and then adapt these results to a more realistic setup where resources are integer, by using rounding heuristics discussed in Section IV-B. Rounding heuristic is a well-known practice commonly used in resource scheduling [10], [27] to round to integer values the rational values assigned to integer variables, which are number of nodes and cores in our case.

A. Problems

We consider two problems in this research, the first problem, CO-SCHED, consists into finding a mapping that minimized the makespan of an ensemble of workflows given that we have a resource allocation, so for each simulation and analysis in the workflow we know on how many nodes/cores they are running. The second problem, CO-ALLOC, is the reverse problem, from a given mapping we aim to compute an optimal resource allocation scheme.

Problem 1 (CO-SCHED). *Given a number nodes and cores assigned to simulations and analysis, find a co-scheduling mapping m^* minimizing the makespan of the entire ensemble.*

Problem 2 (CO-ALLOC). *Given a co-scheduling mapping m , compute the amount of resources $n_{S_i}^*, c_{S_i}^*$ assigned for each simulation $S_i \in S$ and $n_{A_j}^*, c_{A_j}^*$ for each analysis $A_j \in A$ such that the makespan of the entire ensemble is minimized.*

Ideal scenario. To find a co-scheduling mapping which minimizes the makespan for CO-SCHED, we start with a *ideal* co-scheduling mapping which is defined as a mapping where all analyses are co-scheduled with their coupled simulation. This scenario is denoted as *ideal* as it implies that you have enough compute nodes to accommodate such mapping. We show that the makespan is minimized under an ideal co-scheduling mapping by verifying the correctness of the following theorem:

Theorem 1. *The makespan is minimized if, and only if, each analysis is co-scheduled with its coupled simulation.*

Proof. The proof is given in Appendix A. \square

Based on Theorem 1, we have a solution for CO-SCHED but under the strict assumption that we have enough resources to achieve such ideal mapping (i.e., co-schedule all analyses with their respective coupled simulations). However, due to resource constraints, for instance, there is insufficient memory to accommodate all the analyses and their coupled simulation, or bandwidth congestion as multiple analyses performs I/O at the same time, that ideal co-scheduling mapping might be unachievable, and some analyses are not able to co-schedule with their coupled simulation.

Constrained resources scenario. We therefore have to consider co-scheduling mappings in which there are analyses that are not co-scheduled with their coupled simulation. There are numerous such co-scheduling mappings to explore, indeed each analysis that are not co-scheduled with its coupled simulation could potentially be co-scheduled with any other applications. To reduce number of considered co-scheduling mappings, we show that the makespan is minimized if, and only if, analyses that are not co-scheduled with their coupled simulations are co-scheduled inside analysis-only co-scheduling allocations, i.e. without the presence of any simulation. Based on this observation, we build a partition of analyses which are co-scheduled with their respective simulations and co-scheduled with other analyses (see Figure 1).

Theorem 2. *Given a set of analyses that are not co-scheduled with their coupled simulation, the makespan of the ensemble is minimized when these analyses are co-scheduled within analysis-only co-scheduling allocations.*

Proof. The proof is given in Appendix B. \square

Note that, since analyses co-scheduled on analysis-only allocations have no data dependency among each other (they only communicate with their simulations), adjusting co-scheduling placements among them has no impact on the makespan.

B. Resource Allocation for CO-ALLOC

1) *Rational Allocation:* We first show how to compute the (rational) numbers of nodes and cores assigned to each simulation and analysis in the workflow ensemble. In other words, we demonstrate how to solve CO-ALLOC. The idea is to allocate resources to co-scheduled applications such that differences among execution time of every co-scheduling allocations are minimized, thereby leading to minimal makespan. The intuition is that if one allocation has a smaller execution time than another allocation, we can take resources (thanks to rational number of nodes and cores) from the faster allocation to accelerate the slower one until all allocations finish approximately at the same time, hence improving the overall makespan.

Analysis-only allocations. Based on the above reasoning, we first compute resource allocation for analysis-only co-scheduling allocations (c.f. Theorem 3). Let P^{NC} denote the set of analyses that are not co-scheduled with their coupled simulation. Based on Theorem 2, we have to find a resource allocation for the co-scheduling mapping where analyses in P^{NC} are co-scheduled on analysis-only co-scheduling allocations. Let assume that the analysis in P^{NC} are distributed among L analysis-only co-scheduling allocations, in which each allocation is a no-intersecting subset of analyses denoted by P_i^{NC} . Given X is a set of jobs, which are simulations or analyses, for the ease of presentation, we define the following notations:

- $T(X) = \max_{x \in X} t(x)$ is the time to execute one iteration of X concurrently;
- $Q(X) = \sum_{x \in X} t_x(1)$ is the time to execute sequentially one iteration of X , i.e. executing on single core;
- $U(X) = \sum_{x \in X} c_x V_x$ is the cost of processing data remotely for X .

Theorem 3. *To minimize makespan, number of nodes and cores assigned to each analysis in analysis-only co-scheduling allocation $\langle P_i^{NC} \rangle$ are as follows:*

$$n_{\langle P_i^{NC} \rangle}^* = \frac{BQ(P_i^{NC}) + U(P_i^{NC})}{BQ(S \cup A) + U(P^{NC})} N \quad (4)$$

$$c_{A_k}^* = \frac{BQ(A_k)}{BQ(P_i^{NC}) + U(P_i^{NC}) - CV_{A_k}} C, \forall A_k \in P_i^{NC} \quad (5)$$

Proof. The proof is given in Appendix C. \square

Now the remaining part is to find $U(P_i^{NC})$. To prevent from underutilized resources, for each analysis-only co-scheduling allocation containing P_i^{NC} , we expect $\sum_{A_k \in P_i^{NC}} c_{A_k} = C$. By substituting Equation (5) to this equality, we have:

$$\sum_{A_k \in P_i^{NC}} \frac{Q(A_k)}{BQ(P_i^{NC}) + U(P_i^{NC}) - CV_{A_k}} = \frac{1}{B} \quad (6)$$

Because the left-hand side of Equation (6) is strictly decreasing when $U(P_i^{NC})$ is increased, Equation (6) has only a real root $U^*(P_i^{NC})$. In the experiments, we use SymPy, a Python package for symbolic computing to solve it numerically. Substituting $U^*(P_i^{NC})$ to Equations (4) and (5), resource assignment for the analysis-only co-scheduling allocations is determined.

Simulation-based Allocations. In this second step, we find a resource allocation for the remaining co-scheduling allocations, in which every simulation is co-scheduled with a subset of analyses it couples with. Specifically, we distribute remaining resources of $N - \tilde{N}$ nodes, where $\tilde{N} = \sum_{i=1}^L n_{\langle P_i^{NC} \rangle}^*$ (the value of $n_{\langle P_i^{NC} \rangle}^*$ is computed from Theorem 3), for each simulation $S_i \in S$ and each analysis $A_j \in A \setminus P^{NC}$. Recall that $m(S_i)$ is the set of analyses that are co-scheduled with S_i .

Theorem 4. To minimize makespan, number of nodes and cores assigned to each simulation and analysis co-scheduled on co-scheduling allocation $\langle S_i \cup m(S_i) \rangle$ are as follows:

$$n_{S_i}^* = \frac{Q(S_i \cup m(S_i))}{Q(S \cup A \setminus P^{NC})} (N - \tilde{N}) \quad (7)$$

$$c_{S_i}^* = \frac{Q(S_i)}{Q(S_i \cup m(S_i))} C \quad (8)$$

$$c_{A_j}^* = \frac{Q(A_j)}{Q(S_i \cup m(S_i))} C, \forall A_j \in m(S_i) \quad (9)$$

Proof. The proof is given in Appendix D. \square

2) *Integer Allocation:* We have a solution to CO-ALLOC with rational numbers of resources, however a more practical solution must use integer number of cores and nodes. Therefore, we relax the rational solution to an integer solution by applying a resource-preserving rounding heuristic. There are two levels of rounding needed to be handled: node-level rounding and core-level rounding. The node-level one rounds the number of nodes assigned to each co-scheduling allocation, while the core-level rounds the number of cores per node assigned to each application co-scheduled within the same co-scheduling allocation. The objective is to keep the makespan as minimized as possible after rounding.

These rounding problems are all declared as *sum preserving rounding* as the sum of nodes or cores per node are the same after rounding. Formally, let assume we would like to round x_i to an integer value x_i^I such that $\sum x_i^I = \sum x_i = s$. For every x_i , a rounding way is to determine $i(x_i) \in \{0, 1\}$ such that $x_i^I = x_i + i(x_i)$. Hence, $\sum i(x_i) = s - \sum \lfloor x_i \rfloor$. The idea is to pick $s - \sum \lfloor x_i \rfloor$ numbers among x_i and round them up while round the others down, so that the sum of them are preserved.

The remaining task is to determine the number of nodes or cores of which jobs to round up based on the aforementioned objective of minimizing makespan, or minimizing the difference in new T , i.e. time to execute one iteration, after rounding (see Equation (3)). From Equations (1) and (2), since assigned resources (i.e. number of nodes, number of cores per node) are inversely proportional to time to execution one iteration, one core, we present a rounding heuristic based on simulation execution time. Specifically, at core-level, among applications x in a co-scheduling allocation, we prioritize to round up whose $t_x(1)$ is larger. At node-level, among co-scheduling allocations X , we prioritize to round up $n_{\langle X \rangle}$ whose $t_{\langle X \rangle}$ is larger. The algorithm following this rounding heuristic is described in Algorithm 1.

V. EVALUATION

In this section, we present an evaluation of our proposed co-scheduling strategies developed to solve CO-SCHED and show the effectiveness of computing resource allocation approach developed to solve CO-ALLOC.

Simulator. We implement a simulator based on WRENCH [28] and SimGrid [29] to accurately emulate the execution of simulations and in situ analyses in a

Algorithm 1: Rounding algorithm at node-level

Input : $N, \tilde{N}, n_{S_i}^*, Q(S_i \cup m^*(S_i))$
Output: $n_{S_i}^I$

```

1  $k \leftarrow N - \tilde{N}$ 
2  $dict \leftarrow \{\}$ 
3 for  $S_i \in S$  do
4    $k \leftarrow k - \lfloor n_{S_i}^* \rfloor$ 
5    $dict \leftarrow dict + \{S_i : Q(S_i \cup m^*(S_i))\}$ 
6 end
7  $sorted\_dict \leftarrow sortByValue(dict)$ 
8  $j \leftarrow 1$ 
9 for  $S_i \in sorted\_dict$  do
10  if  $j > k$  then
11     $n_{S_i}^I \leftarrow \lfloor n_{S_i}^* \rfloor$ 
12  else
13     $n_{S_i}^I \leftarrow \lceil n_{S_i}^* \rceil$ 
14  end
15   $j \leftarrow j + 1$ 
16 end
17 return  $n_{S_i}^I$ 
```

workflow ensemble (see Figure 2). To emulate behavior of simulations and analyses, every simulation and analysis step is divided into two fine-grained stages: computational stage to perform certain simulation or analysis task, and I/O stage to publish/subscribe data to/from the simulation/analysis, respectively. The executions of these fine-grained stages tightly depend on each other following a validated in situ execution model [30] to reflect coupling behavior between the simulation and in situ analyses. Task dependencies serve as an input for simulator workloads to simulate coupling behavior between simulations and in situ analyses. Since a job in WRENCH is simulated on a single node, we emulate multi-node jobs by replicating stages of the job on multiple nodes and extrapolating task dependencies such that the

Simulator (Real, not numerical simulation!!!)

<https://github.com/Analytics4MD/A4MD-insitu-ensemble>

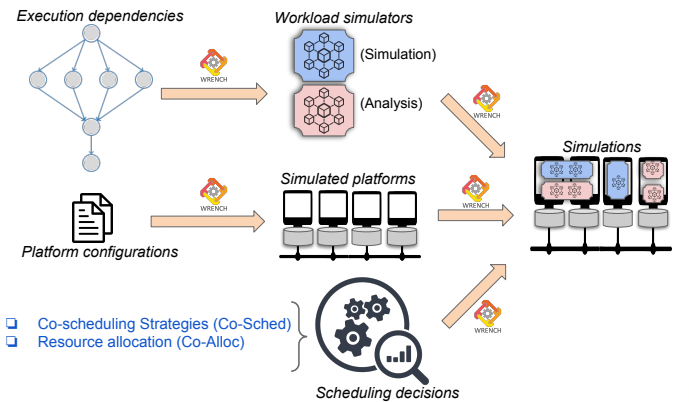


Fig. 2: Simulator

¹<https://github.com/Analytics4MD/insitu-ensemble-simulator>

A. Experimental Setup

Setup. The simulator takes the following inputs: (1) a workflow ensemble’s configurations (i.e., number of simulations, number of coupled analyses per simulation, coupling relations), (2) application profiles (i.e., simulations and analyses’ sequential execution times), (3) resources constraints (i.e., number of nodes, number of cores per node, processor speed, network bandwidth), (3) a co-scheduling mapping (i.e., the simulation each analysis is co-scheduled with), and (4) a resource allocation (i.e., number of nodes and cores each application is assigned). We vary the workflow ensemble’s configurations to evaluate the importance and the impact of different configurations on the efficiency of our proposed scheduling choices. We fix the simulations’ profiles, e.g. sequential execution time of the simulations, to study the impact of analyses on scheduling decisions. Sequential execution time of the analyses is generated randomly in a specific range which is relative to the sequential execution time of their coupled simulations. In this experiment, we generate it from 50% to 150% of the simulation’s sequential execution time. Even though our solution is applicable to any data size generated by the simulations, for the sake of simplicity, the amount of data processed by each analysis is kept at a fixed amount among all simulations, even though that amount is varied at different workflow ensemble’s configurations.

Scenarios. We compare a variety of scenarios corresponding to different co-scheduling mappings (see Table III). IDEAL is the co-scheduling mapping where all analyses are co-scheduled with their respective simulation. IN-TRANSIT represents the scenario where all analyses are co-scheduled together on a dedicated co-scheduling allocation. We also compose hybrid scenarios where some analyses are co-scheduled with the simulation they couple with while the others are co-scheduled on the co-scheduling allocation without the presence of any simulation. For INCREASING- $x\%$ (resp. DECREASING- $x\%$), we pick the largest $x\%$ of the analyses, which is sorted in ascending (resp. descending) order of sequential execution time, to not co-schedule with the simulation they couple with (i.e. to co-schedule them on the dedicated co-scheduling allocation). In this evaluation, we consider different percentages of number of analyses, 25%, 50%, 75%. For each of these co-scheduling mappings, we compute the resource allocations for the ensemble using CO-ALLOC, which is described in Section IV-B.

B. Bandwidth Calibration

In order for the simulator to accurately reflect the execution platform, we have to calibrate the bandwidth per node used in our model to improve the precision of our solution using CO-ALLOC. In our model, the bandwidth model is simple and optimistic as it does not account for concurrent accesses (i.e., the bandwidth available to each application is the maximum bandwidth B). The idea is to compare the makespan given by our performance model (i.e., Theorem 3) to the makespan given by our simulation as WRENCH/SimGrid offers pretty complex and accurate communication models. In Figure 3,

TABLE III: Experimental scenarios, where $A(x\%)$ is the $x\%$ largest analyses’ sequential time of A .

Scenarios	Analyses co-scheduled with their coupled simulation (P^C)	Analyses not co-scheduled with their coupled simulation (P^{NC})
IDEAL	A	\emptyset
IN-TRANSIT	\emptyset	A
INCREASING- $x\%$	$A \setminus A(x\%)$	$A(x\%)$
DECREASING- $x\%$	$A(100 - x\%)$	$A \setminus A(100 - x\%)$

the theoretical makespan estimated by our model using the maximum bandwidth B exhibits a large difference from the makespan resulted by the simulator (up to 30 times). The simulated bandwidth is smaller than B during the execution as it is shared among concurrent I/O operations. Therefore, we explore different bandwidth values (see Table II) and re-estimate the makespan using our model based on the resource allocation at maximum bandwidth.

Model	Bandwidth
Baseline	B
B'_1	$B / P^{NC} $
B'_2	B / \tilde{N}
B'_3	$B / (\tilde{N} \times P^{NC})$

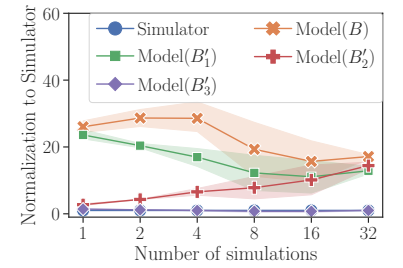


TABLE II: Bandwidth models to calibrate the simulator. \tilde{N} is number of nodes assigned to analysis in P^{NC} .

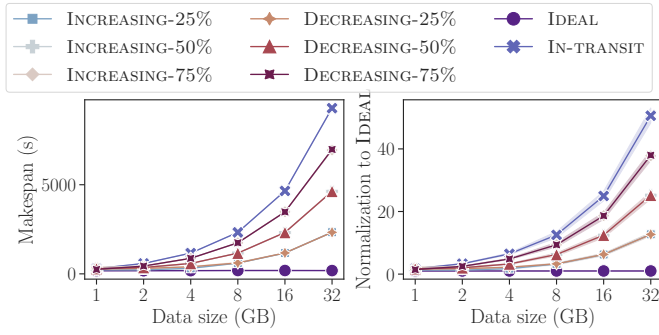
Fig. 3: Makespan estimated by the model for the INCREASING-50% scenario with different values of bandwidth per node (see Table II).

Figure 3 confirms our hypothesis that bandwidth is shared among concurrent I/Os and the accuracy of CO-ALLOC depends on choosing appropriate bandwidth used in the model. B'_3 provides an approximation that is close to the makespan given by the simulator, we therefore choose B'_3 as the calibrated bandwidth for the experiments hereafter.

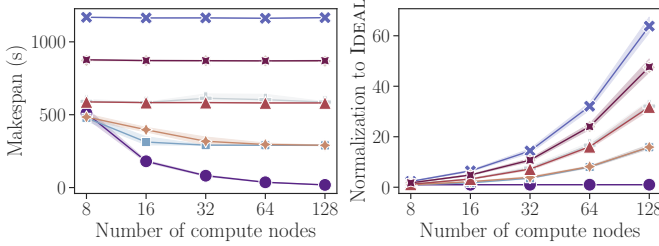
C. Results

In this subsection, we explore the configuration space of the workflow ensemble. We aim to characterize the impact of different co-scheduling strategies by comparing the makespan of the simulator over various scenarios described in Table III. To ensure the reliability of results, each value is averaged over 5 trials.

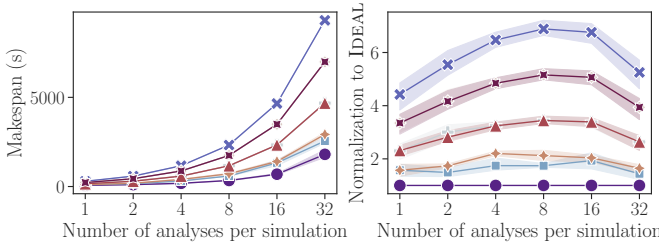
Implications of CO-SCHED. We vary one configuration of the workflow ensemble at a time while keeping the other configurations fixed. Figure 4a shows the makespan of the workflow ensemble running on 16 nodes with 4 simulations, 4 analyses per simulation when varying the size of data processed by each analysis each iteration. Figure 4b is the result with 4 simulations, 4 analyses per simulation and 4GB of data when varying the number of nodes. In Figure 4c, the number of analyses per simulation is varied while keeping the number of simulation fixed at 4 and 4GB of data, 4 analyses per simulation. In Figure 4d, the number of simulations is



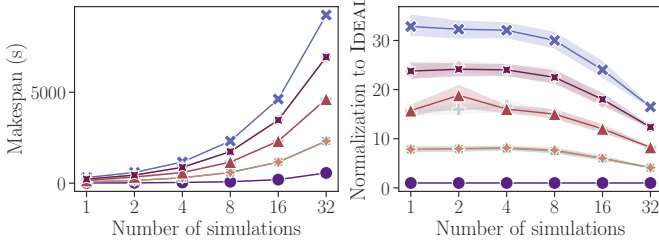
(a) Data size is varying. The workflow ensemble runs on 16 nodes and has 4 simulations, 4 analyses per simulation.



(b) Number of nodes is varying. The workflow ensemble has 4 simulations, 4 analyses per simulation. Each analysis processes 4GB of data each iteration



(c) Number of analyses per simulation is varying. The workflow ensemble runs on 16 nodes and has 4 simulations. Each analysis processes 4GB of data each iteration



(d) Number of simulations is varying. The workflow ensemble runs on 64 nodes and has 4 analyses per simulation. Each analysis processes 4GB of data each iteration

Fig. 4: Makespan of the workflow ensemble when varying various configurations.

varied while keeping the number of analyses per simulation fixed at 4, 4GB of data and 64 nodes.

Figure 4 demonstrates the following order if the scenarios are sorted in ascending order by their makespan: IDEAL, X-25%, X-50%, X-75%, IN-TRANSIT, where X is either INCREASING or DECREASING as those scenarios' makespan are approximately close to each other. Note that, for IDEAL, 0% of the analyses are not co-scheduled with their coupled

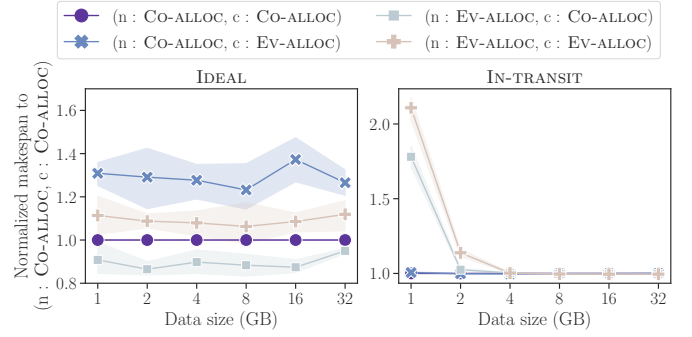


Fig. 5: Comparison of normalized makespan of CO-ALLOC to EV-ALLOC, where resources are evenly distributed. (n : CO-ALLOC, c : EV-ALLOC) indicates CO-ALLOC is applied at node-level while EV-ALLOC is applied at core-level.

simulation while that percentage is 100% for IN-TRANSIT. This observation aligns with our finding in Theorem 1 that not co-scheduling an analysis with its coupled simulation yields a higher makespan. The more number of analyses not co-scheduled with their coupled simulation, the slower the makespan. Since IDEAL outperforms the other scenarios, ideal co-scheduling mappings should be favored when the available resources can sustain. In Figures 4a, 4c and 4d, the makespan scales linearly with the size of data processed by the analyses each iteration, the number of analyses per simulation and the number of simulations, respectively in which IDEAL imposes the smallest escalation when these parameters are increased. In Figure 4b, only the makespan of the IDEAL is decreased when the number of nodes is increased. This is because the more nodes are utilized, the more communications are required to exchange the data, which requires the communication bandwidth is shared among them. The communication stages in each step are therefore slower even though the computational stages are faster due to more computing resources are assigned. Only the IDEAL which incurs no remote communication can fully take advantage of the available resources.

Efficiency of CO-ALLOC. We compare the resource allocation computed by CO-ALLOC with a naive approach of evenly distributing the resources, which is called as EV-ALLOC. Specifically, at node-level, with EV-ALLOC, the number of nodes is equally divided among co-scheduling allocations while at core-level, for each co-scheduling allocation, the number of cores per node is equally divided among applications co-scheduled in such the allocation. Let $(n : x, c : y)$ is a manner to compute resource allocations, in which we apply x method at node-level while using y method at core-level. We study 4 possible combinations: $(n : \text{CO-ALLOC}, c : \text{CO-ALLOC})$, $(n : \text{CO-ALLOC}, c : \text{EV-ALLOC})$, $(n : \text{EV-ALLOC}, c : \text{CO-ALLOC})$, and $(n : \text{EV-ALLOC}, c : \text{EV-ALLOC})$.

Even though the experiment is conducted over all the scenarios in the configuration space, due to the lack of space, we selectively present a subset of scenarios where the allocations significantly differ between observing combina-

tions. Figure 5 shows the makespan, which is normalized to $(n : \text{CO-ALLOC}, c : \text{CO-ALLOC})$, in two representative extreme scenarios: IDEAL and IN-TRANSIT when the amount of data processed each iteration is varied. In the IDEAL scenario, $(n : \text{CO-ALLOC}, c : \text{CO-ALLOC})$ surpasses the other combinations, except $(n : \text{EV-ALLOC}, c : \text{CO-ALLOC})$. This is because IDEAL comprises no remote communication, thus execution time of each application's step is dominant by the portion of computational stages, which cause the performance is more sensitive to the changes at core-level than at node-level. For IN-TRANSIT, using EV-ALLOC at node-level results in a slower makespan (up to two times higher) than using CO-ALLOC as off-node communications are primary in this scenario. For larger data sizes, the I/O stages take the large proportion in each step in IN-TRANSIT, which makes the computing time negligible, thus there is no clear difference among discussed combinations.

VI. CONCLUSIONS

In this paper, we have introduced an execution model of coupling behavior between in situ jobs in a complex workflow ensemble. Based on the proposed model, we theoretically characterized computation and communication characteristics in each in situ component. We further relied on this model to determine co-scheduling policies and resource profiles for each simulation and in situ analysis such that the makespan of the workflow ensemble is minimized under given available computing resources. By evaluating the scheduling solutions via simulation, we validate our proposed approach and confirm its correctness. We are also able to confirm the relevance of data locality as well as the need of well-management shared resources (e.g., bandwidth per node) in co-scheduling concurrent applications together.

For future work, we plan to extend the scheduling model to include contention and interference effects between in situ jobs sharing the same underlying resources. Another promising research direction is to consider more complex use cases leveraging ensembles, e.g. adaptive sampling [31] in which more coordination will be needed to achieve better performance as the simulations are periodically stopped and restarted.

ACKNOWLEDGMENTS

This work is funded by NSF contracts #1741040, #1741057, and #1841758.

APPENDIX

Let X is a set of jobs, which are simulations or analyses. We define the following:

- $T(X) = \max_{x \in X} t(x)$
- $Q(X) = \sum_{x \in X} t_x(1)$
- $U(X) = \sum_{x \in X} c_x V_x$

Remark 1. To minimize $T(X) = \max_{x \in X} t(x)$ given that resources are finite, we allocate resources such that every $t(x)$, where $x \in X$, is equal to minimize difference among $t(x)$.

We use this remark to minimize execution time of concurrent applications. This remark is held when the number of resources (i.e. number of compute nodes, number of cores) are rational numbers. We also note that all the following proofs are given under the context of rational allocation. We discuss how to relax to integer allocations in Section IV-B2.

Based on Remark 1, we compute minimal execution time in one iteration of a co-scheduling allocation, which is foundation for our following proofs.

Lemma 1. Given a co-scheduling allocation $\langle S_i, m(S_i) \rangle$, and P_i is the set of analyses that are co-scheduled with S_i , but not coupled with S_i . The execution time for one iteration is minimized at:

$$T^*(S_i \cup m(S_i)) = \frac{BQ(S_i \cup m(S_i)) + U(P_i)}{BCn_{S_i}} \quad (10)$$

Proof of Lemma 1. For every analysis $A_j \in m(S_i) \setminus P_i$ (set of analyses co-scheduled with S_i and coupled with S_i) and $A_k \in P_i$ (set of analyses co-scheduled with S_i , but not coupled with S_i), $T(S_i \cup m(S_i))$ is minimized at $T^*(S_i \cup m(S_i))$ when:

$$T^*(S_i \cup m(S_i)) = t(S_i) = t(A_j) = t(A_k) \quad (11)$$

Since $A_k \in P_i$ is not coupled with S_i , from Equations (1) and (2), we have:

$$\begin{aligned} \frac{t_{S_i}(1)}{n_{S_i} c_{S_i}} &= \frac{t_{A_j}(1)}{n_{A_j} c_{A_j}} = \frac{t_{A_k}(1)}{n_{A_k} c_{A_k}} + \frac{V_{A_k}}{Bn_{A_k}} \\ &= \frac{BQ(S_i \cup m(S_i)) + U(P_i)}{B(n_{S_i} c_{S_i} + \sum_{A_j \in m(S_i)} n_{A_j} c_{A_j})} \end{aligned} \quad (12)$$

We note that $n_{S_i} = n_{A_j}$, as they are co-scheduled on the same co-scheduling allocation. To avoid resources are underutilized, we expect the entire cores of every allocation are used, or:

$$c_{S_i} + \sum_{A_j \in m(S_i)} c_{A_j} = C \quad (13)$$

Substituting Equation (13) to Equation (12), we obtain Equation (10), which is what we need to prove. \square

A. Proof of Theorem 1

We compare the minimized makespan of two following co-scheduling mappings: (1) the original co-scheduling mapping m and (2) the co-scheduling mapping m' forming from m by not co-scheduling an analysis A_k with its coupled simulation. We are going to prove that the makespan of m' is greater than one of m , which implies not co-scheduling an analysis with the its coupled simulation causes the makespan slower. We therefore, indirectly prove the theorem.

Since A_k is co-scheduled with its coupled simulation in the co-scheduling mapping m , from Remark 1 and Equation (3), $T(S \cup A)$ of the co-scheduling mapping m , is minimized when:

$$T(S \cup A) = t(A_k) = \frac{t_{A_k}(1)}{n_{A_k} c_{A_k}} \quad (14)$$

On the other hand, in the co-scheduling mapping m' , A_k is not co-scheduled with its coupled simulation, then $T'(S \cup A)$ is minimized when:

$$T'(S \cup A) = t'(A_k) = \frac{t_{A_k}(1)}{n'_{A_k} c'_{A_k}} + \frac{V_{A_k}}{B n'_{A_k}} \quad (15)$$

Proof by Contradiction. Let assume $T(S \cup A) \geq T'(S \cup A)$, then $t(A_k) \geq t'(A_k)$. From Equations (14) and (15), we have:

$$\frac{t_{A_k}(1)}{n_{A_k} c_{A_k}} \geq \frac{t_{A_k}(1)}{n'_{A_k} c'_{A_k}} + \frac{V_{A_k}}{B n'_{A_k}} > \frac{t_{A_k}(1)}{n'_{A_k} c'_{A_k}} \quad (16)$$

$$n_{A_k} c_{A_k} < n'_{A_k} c'_{A_k}$$

Let $r(x) = n_x c_x$ denote the amount of resources in terms of total cores assigned to x . Equation (16) indicates more resources assigned to A_k in m' than m . However, as the total number of cores is limited to $N \times C$, in the co-scheduling m' , there must exist an application I which is either a simulation in S or an analysis in $A \setminus A_k$ such that it is assigned fewer resources than it is in the co-scheduling m . Hence, $T'(I) > T(I)$, which makes $T(S \cup A) < T'(S \cup A)$. This is contradicted to our assumption. Hence, from Equation (3), the makespan of co-scheduling m is smaller than the makespan of m' .

B. Proof of Theorem 2

We compare the minimized makespan of two following co-scheduling mappings: (1) the co-scheduling mapping m^* where all analyses not co-scheduled with their coupled simulation, denoted by P^{NC} , are co-scheduled on analysis-only co-scheduling allocations and (2) the co-scheduling mapping m' in which not all analyses in P^{NC} are co-scheduled on analysis-only co-scheduling allocations. Hence, in this mapping m' , P^{NC} is able to partition into two parts, P'^{NC} is the set of analyses that are co-scheduled in analysis-only co-scheduling allocations, and P'^C is the set of analyses that are co-scheduled on the co-scheduling allocations that has the simulations they do not couple with. Then $P'^C \cup P'^{NC} = P^{NC}$. We also denote the set of analyses co-scheduled with their coupled simulation in these co-scheduling mappings by P^C , thus $P^C \cup P^{NC} = A$. We are going to prove the makespan of m' is greater than the makespan of m^* .

Co-scheduling mapping m^* . For the mapping m^* , for every co-scheduling allocation $\langle S_i \cup m^*(S_i) \rangle$, from Lemma 1, $T(S_i \cup m^*(S_i))$ is minimized at:

$$T^*(S_i \cup m^*(S_i)) = \frac{BQ(S_i \cup m^*(S_i))}{BC n_{S_i}^*} \quad (17)$$

Based on Remark 1, $T(S \cup P^C)$ is minimized when every $T^*(S_i \cup m^*(S_i))$ is equal, or

$$T^*(S \cup P^C) = \frac{B \sum_{S_i \in S} Q(S_i \cup m^*(S_i))}{BC \sum_{S_i \in S} n_{S_i}^*} = \frac{BQ(S \cup P^C)}{BC \sum_{S_i \in S} n_{S_i}^*} \quad (18)$$

Following the same procedure for analysis-only co-scheduling allocations in m^* , from Lemma 1, $T(P^{NC})$ is minimized at:

$$T^*(P^{NC}) = \frac{BQ(P^{NC}) + U^*(P^{NC})}{BC \tilde{N}^*}, \quad (19)$$

where \tilde{N}^* is total number of nodes assigned to analysis-only allocations.

Now combining simulation-present co-scheduling allocations and analysis-only co-scheduling allocations, $T(S \cup A)$ is minimized, denoted by t^* , when $t^* = T^*(S \cup P^C) = T^*(P^{NC})$. From Equations (18) and (19):

$$t^* = \frac{BQ(S \cup P^C)}{BC \sum_{S_i \in S} n_{S_i}^*} = \frac{BQ(P^{NC}) + U^*(P^{NC})}{BC \tilde{N}^*} \quad (20)$$

$$t^* = \frac{BQ(S \cup A) + U^*(P^{NC})}{NBC} \quad (21)$$

$$U^*(P^{NC}) = \frac{BQ(S \cup P^C)}{\sum_{S_i \in S} n_{S_i}^*} N - BQ(S \cup A) \quad (21)$$

Co-scheduling mapping m' . Similarly, with the mapping m' , $T(S \cup A)$ is minimized, denoted by t' when $t' = T'(S \cup P^C \cup P'^C) = T'(P'^{NC})$. For the mapping m' , we use prime notations to differentiate from the mapping m^* .

$$t' = \frac{BQ(S \cup P^C \cup P'^C) + U'(P'^C)}{BC \sum_{S_i \in S} n'_{S_i}} = \frac{BQ(P'^{NC}) + U'(P'^{NC})}{BC \tilde{N}'}$$

$$t' = \frac{BQ(S \cup A) + U'(P'^{NC})}{NBC} \quad (22)$$

$$U'(P'^{NC}) = \frac{BQ(S \cup P^C \cup P'^C) + U'(P'^C)}{\sum_{S_i \in S} n'_{S_i}} N - BQ(S \cup A) \quad (23)$$

Compare t' and t^* . For the sake of simplicity, we denote:

- $n' = \sum_{S_i \in S} n'_{S_i}$ and $n^* = \sum_{S_i \in S} n_{S_i}^*$
- $U^* = U^*(P^{NC})$ and $U' = U'(P'^{NC})$

From Equations (20) and (22), we have:

$$t' - t^* = \frac{U'(P'^{NC}) - U^*(P^{NC})}{NBC} \quad (24)$$

Thus to compare t^* and t' , we only need to compare $U^*(P^{NC})$ and $U'(P'^{NC})$.

Moreover, from Equations (21) and (23):

$$U' - U^* = \frac{N}{n'} \left(\frac{BQ(S \cup P^C)}{n^*} (n^* - n') + U'(P'^C) + BQ(P'^C) \right) \quad (25)$$

Again, from Lemma 1, we have:

$$t^* = \frac{BQ(S \cup P^C)}{BC n^*} \Rightarrow \frac{BQ(S \cup P^C)}{n^*} = t^* BC \quad (26)$$

Substituting Equation (26) to Equation (25):

$$U' - U^* = \frac{N}{n'} (t^* BC (n^* - n') + U'(P'^C) + BQ(P'^C)) \quad (27)$$

Let us define the set of simulations that analyses in P'^C co-scheduled with as S'^C . We note that $|S'^C| \leq |P'^C|$, then again from Lemma 1:

$$t' = T'(S'^C \cup P'^C) = \frac{BQ(P'^C) + U'(P'^C)}{BC \sum_{S_i \in S'^C} n_{S_i}} \quad (28)$$

By substituting Equation (24) to Equation (28):

$$BQ(P'^C) + U'(P'^C) = (t^* BC + \frac{U' - U^*}{N}) \sum_{S_i \in S'^C} n'_{S_i} \quad (29)$$

By substituting Equation (29) to Equation (27), we have:

$$(U' - U^*)(n' - \sum_{S_i \in S'^C} n'_{S_i}) = Nt^* BC(n^* + \sum_{S_i \in S'^C} n'_{S_i} - n') \quad (30)$$

Since the total number of compute nodes for co-scheduling S and P^C plus the number of nodes of co-scheduling allocations containing analyses in P'^C is always greater than the total number of nodes for co-scheduling S , P^C and P'^C , then

$$n^*_{\langle S, P^C \rangle} + \sum_{S_i \in S'^C} n'_{\langle S_i, m'(S_i) \rangle} > n'_{\langle S, P^C, P'^C \rangle} \quad (31)$$

$$n^* + \sum_{S_i \in S'^C} n'_{S_i} > n' \quad (32)$$

We also note that as $S'^C \subseteq S$ then:

$$n' = \sum_{S_i \in S} n'_{S_i} > \sum_{S_i \in S'^C} n'_{S_i} \quad (32)$$

From Equations (30) to (32), we have $U' > U^*$, or $t' > t^*$.

C. Proof of Theorem 3

Let \tilde{N} denote total number of nodes assigned to analysis-only co-scheduling allocations. Hence, $\tilde{N} = \sum_{i=1}^L n^*_{\langle P_i^{NC} \rangle}$, where $\cup_{i=1}^L P_i^{NC} = P^{NC}$. We determine resource assignment to each analysis-only co-scheduling allocation at node- and core-level, respectively.

Node-level Allocation. To minimize makespan of entire ensemble (see Equation (3)), we need to minimize $T(S \cup A)$ by verifying that T of analysis-only co-scheduling allocations is equal to T of simulation-based co-scheduling allocations (according to Remark 1), or $T^*(S \cup P^C) = T^*(P^{NC})$. By applying Lemma 1:

$$\frac{Q(S \cup P^C)}{C(N - \tilde{N})} = \frac{BQ(P^{NC}) + U(P^{NC})}{BC\tilde{N}} \quad (33)$$

$$\tilde{N} = \frac{BQ(P^{NC}) + U(P^{NC})}{BQ(S \cup A) + U(P^{NC})} N$$

Among analysis-only co-scheduling allocations $\langle P_i^{NC} \rangle$, from Remark 1, to minimize $T(P^{NC})$, we verify $T^*(P^{NC}) = T^*(P_i^{NC})$, $\forall i \in \{1, \dots, L\}$. Similarly, applying Lemma 1:

$$\frac{BQ(P^{NC}) + U(P^{NC})}{BC\tilde{N}} = \frac{BQ(P_i^{NC}) + U(P_i^{NC})}{BCn^*_{\langle P_i^{NC} \rangle}} \quad (34)$$

Substituting Equation (33) to Equation (34), the number of nodes assigned to each analysis-only co-scheduling allocation is computed as:

$$n^*_{\langle P_i^{NC} \rangle} = \frac{BQ(P_i^{NC}) + U(P_i^{NC})}{BQ(S \cup A) + U(P^{NC})} N$$

Core-level Allocation. Similarly, based on Remark 1, for each analysis-only co-scheduling allocation $\langle P_i^{NC} \rangle$, $T(P_i^{NC})$ is minimized when $T^*(P_i^{NC}) = T^*(A_k)$ for every $A_k \in P_i^{NC}$. From Equations (2) and (34):

$$\frac{BQ(P_i^{NC}) + U(P_i^{NC})}{BCn^*_{\langle P_i^{NC} \rangle}} = \frac{t_{A_k}(1)}{n^*_{A_k} c_{A_k}} + \frac{V_{A_k}}{Bn^*_{A_k}} \quad (35)$$

Note that $n^*_{\langle P_i^{NC} \rangle} = n^*_{A_k}$ as A_k is co-scheduled on $\langle P_i^{NC} \rangle$. Hence, from Equation (35), the number of cores assigned to each analysis in analysis-only co-scheduling allocation expressed as:

$$c^*_{A_k} = \frac{BQ(A_k)}{BQ(P_i^{NC}) + U(P_i^{NC}) - CV_{A_k}} C$$

D. Proof of Theorem 4

Similarly to the method used to prove Theorem 3, we solve the number of nodes and cores per node assigned to each application in simulation-based co-scheduling allocations, respectively.

Node-level Allocation. Remark 1 shows that $T(S \cup P^C)$ is minimized when $T^*(S \cup P^C) = T^*(S_i \cup m(S_i))$ for every $S_i \in S$. From Lemma 1, the number of nodes assigned to each simulation-based co-scheduling allocation is computed $\langle S_i \cup m(S_i) \rangle$ as follows:

$$\frac{Q(S \cup P^C)}{C(N - \tilde{N})} = \frac{Q(S_i \cup m(S_i))}{Cn^*_{\langle S_i \cup m(S_i) \rangle}} \quad (36)$$

$$n^*_{\langle S_i \cup m(S_i) \rangle} = \frac{Q(S_i \cup m(S_i))}{Q(S \cup P^C)} (N - \tilde{N})$$

Core-level Allocation. For each co-scheduling allocation, $T(S_i \cup m(S_i))$ is minimized when $T^*(S_i \cup m(S_i)) = T^*(S_i)$. Applying Lemma 1:

$$\frac{Q(S_i \cup m(S_i))}{Cn^*_{\langle S_i \cup m(S_i) \rangle}} = \frac{Q(S_i)}{n^*_{S_i} c^*_{S_i}} \quad (37)$$

As S_i is co-scheduled in $\langle S_i \cup m(S_i) \rangle$, thus $n^*_{\langle S_i \cup m(S_i) \rangle} = n^*_{S_i}$. From Equation (37), the number of cores per node assigned to each simulation S_i co-scheduled in $\langle S_i \cup m(S_i) \rangle$:

$$c^*_{S_i} = \frac{Q(S_i)}{Q(S_i \cup m(S_i))} C$$

For every analysis $A_j \in m(S_i)$ is co-scheduled on co-scheduling allocation $\langle S_i \cup m(S_i) \rangle$, $T(S_i \cup m(S_i))$ is minimized when $T^*(S_i \cup m(S_i)) = T^*(A_j)$. From Lemma 1,

the number of cores per node assigned to each analysis A_j co-scheduled in $\langle S_i \cup m(S_i) \rangle$:

$$\frac{Q(S_i \cup m(S_i))}{C n_{\langle S_i \cup m(S_i) \rangle}^*} = \frac{Q(A_j)}{n_{A_j}^* c_{A_j}^*}$$

$$c_{A_j}^* = \frac{Q(A_j)}{Q(S_i \cup m(S_i))} C$$

REFERENCES

- [1] I. Foster *et al.*, “Computing just what you need: Online data analysis and reduction at extreme scales,” in *Euro-Par 2017: Parallel Processing*. Cham: Springer International Publishing, 2017, pp. 3–19.
- [2] P. Raiteri *et al.*, “Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics,” *The Journal of Physical Chemistry B*, vol. 110, no. 8, pp. 3533–3539, 2006.
- [3] J. Comer *et al.*, “Multiple-replica strategies for free-energy calculations in namd: Multiple-walker adaptive biasing force and walker selection rules,” *Journal of Chemical Theory and Computation*, vol. 10, no. 12, 2014.
- [4] Y. Okamoto, “Generalized-ensemble algorithms: enhanced sampling techniques for monte carlo and molecular dynamics simulations,” *Journal of Molecular Graphics and Modelling*, vol. 22, no. 5, 2004.
- [5] R. Chelli *et al.*, “Serial generalized ensemble simulations of biomolecules with self-consistent determination of weights,” *Journal of Chemical Theory and Computation*, vol. 8, no. 3, 2012.
- [6] M. Malawski *et al.*, “Algorithms for cost- and deadline-constrained provisioning for scientific workflow ensembles in iaas clouds,” *Future Generation Computer Systems*, vol. 48, pp. 1–18, 2015.
- [7] D. H. Ahn *et al.*, “Flux: Overcoming scheduling challenges for exascale workflows,” *Future Generation Computer Systems*, vol. 110, pp. 202–213, 2020.
- [8] J. L. Peterson *et al.*, “Enabling machine learning-ready hpc ensembles with merlin,” *Future Generation Computer Systems*, 2022.
- [9] P. Malakar *et al.*, “Optimal execution of co-analysis for large-scale molecular dynamics simulations,” in *SC ’16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2016, pp. 702–715.
- [10] G. Aupy *et al.*, “Modeling high-throughput applications for in situ analytics,” *The International Journal of High Performance Computing Applications*, vol. 33, no. 6, pp. 1185–1200, 2022/03/01 2019.
- [11] M. Taufer *et al.*, “Characterizing in situ and in transit analytics of molecular dynamics simulations for next-generation supercomputers,” in *2019 15th International Conference on eScience (eScience)*, 2019, pp. 188–198.
- [12] J. Breitbart *et al.*, “Dynamic co-scheduling driven by main memory bandwidth utilization,” in *2017 IEEE International Conference on Cluster Computing (CLUSTER)*, 2017, pp. 400–409.
- [13] F. V. Zacarias *et al.*, “Intelligent colocation of workloads for enhanced server efficiency,” in *2019 31st International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, 2019, pp. 120–127.
- [14] R. Kuchumov *et al.*, “Analytical and numerical evaluation of co-scheduling strategies and their application,” *Computers*, vol. 10, no. 10, 2021.
- [15] C. Sewell *et al.*, “Large-scale compute-intensive analysis via a combined in-situ and co-scheduling workflow approach,” in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC ’15. New York, NY, USA: Association for Computing Machinery, 2015.
- [16] G. Aupy *et al.*, “Co-scheduling hpc workloads on cache-partitioned cmp platforms,” in *2018 IEEE International Conference on Cluster Computing (CLUSTER)*, 2018, pp. 348–358.
- [17] T. M. A. Do *et al.*, “Performance assessment of ensembles of in situ workflows under resource constraints,” *Concurrency and Computation: Practice and Experience*, vol. n/a, no. n/a, p. e7111, 2022/07/11 2022.
- [18] T. Johnston *et al.*, “In situ data analytics and indexing of protein trajectories,” *Journal of Computational Chemistry*, vol. 38, no. 16, pp. 1419–1430, 2022/07/14 2017.
- [19] D. Dauwe *et al.*, “Modeling the effects on power and performance from memory interference of co-located applications in multicore systems,” in *International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, 1 2014.
- [20] T. M. A. Do *et al.*, “A lightweight method for evaluating in situ workflow efficiency,” *Journal of Computational Science*, vol. 48, p. 101259, 2021.
- [21] G. M. Amdahl, “Validity of the single processor approach to achieving large scale computing capabilities,” in *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*, ser. AFIPS ’67. New York, NY, USA: Association for Computing Machinery, 1967, p. 483–485.
- [22] M. Herlihy *et al.*, *The Art of Multiprocessor Programming, Revised Reprint*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2012.
- [23] T. M. A. Do *et al.*, “Enabling data analytics workflows using node-local storage,” in *International Conference for High Performance Computing, Networking, Storage, and Analysis (SC18)*, Dallas, TX, 2018.
- [24] H. Kim *et al.*, “Understanding energy aspects of processing-near-memory for hpc workloads,” in *Proceedings of the 2015 International Symposium on Memory Systems*, ser. MEMSYS ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 276–282.
- [25] S. Williams *et al.*, “Roofline: An insightful visual performance model for multicore architectures,” *Commun. ACM*, vol. 52, no. 4, p. 65–76, Apr. 2009.
- [26] D. G. Feitelson, “Job scheduling in multiprogrammed parallel systems,” 1997.
- [27] M. Stillwell *et al.*, “Resource allocation algorithms for virtualized service hosting platforms,” *Journal of Parallel and Distributed Computing*, vol. 70, no. 9, pp. 962–974, 2010.
- [28] H. Casanova *et al.*, “Developing Accurate and Scalable Simulators of Production Workflow Management Systems with WRENCH,” *Future Generation Computer Systems*, vol. 112, pp. 162–175, 2020.
- [29] —, “Versatile, scalable, and accurate simulation of distributed applications and platforms,” *Journal of Parallel and Distributed Computing*, vol. 74, no. 10, pp. 2899–2917, 06 2014.
- [30] T. M. A. Do *et al.*, “A lightweight method for evaluating in situ workflow efficiency,” *Journal of Computational Science*, vol. 48, 2021.
- [31] V. Balasubramanian *et al.*, “Adaptive ensemble biomolecular applications at scale,” *SN Computer Science*, vol. 1, no. 2, p. 104, 2020.