ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2018
T. Beeler and N. Thuerey
(Guest Editors)

Volume 37 (2018), Number 8

Time-Domain Parallelization for Accelerating Cloth Simulation

Junbang Liang[1] and Ming C. Lin[1,2]

[1]University of North Carolina at Chapel Hill
[2]University of Maryland at College Park

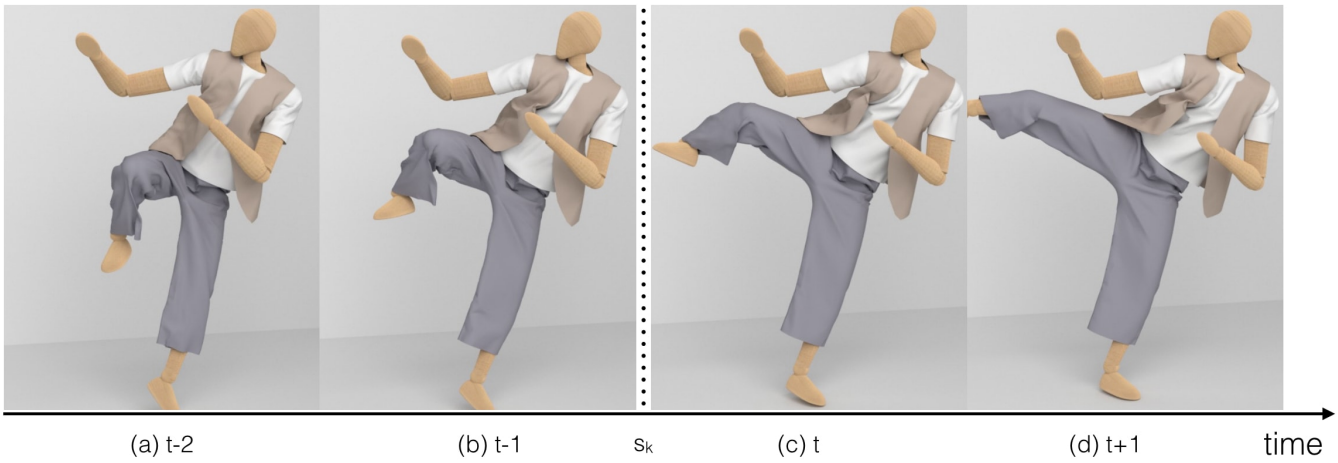


Figure 1: *Simulated 'Karate' animation using our method. Our method parallelizes the simulation workload in time domain using a two-level mesh representation. In the figure, the time domain partition point s_k is between frame t-1 and t, which will be simulated by two different processors. We use an iterative detail recovery algorithm to refine the state of the cloth from low-resolution mesh before the parallel high-resolution simulation begins. As a result, very little visual artifacts can be observed from (b) to (c). In the shown benchmark above, our parallelization method has achieved up to 99x speedup on 128-core systems – an unprecedented level of scalability in distributed CPU systems – compared to at most 47x on a 128-core system [NKT15]. The performance gain is also better than the GPU parallelization [TWT*16] on similar benchmarks, while our approach offers the additional flexibility for coupling with adaptively remeshed cloth simulators.*

Abstract

Cloth simulations, widely used in computer animation and apparel design, can be computationally expensive for real-time applications. Some parallelization techniques have been proposed for visual simulation of cloth using CPU or GPU clusters and often rely on parallelization using spatial domain decomposition techniques that have a large communication overhead. In this paper, we propose a novel time-domain parallelization technique that makes use of the two-level mesh representation to resolve the time-dependency issue and develop a practical algorithm to smooth the state transition from the corresponding coarse to fine meshes. A load estimation and a load balancing technique used in online partitioning are also proposed to maximize the performance acceleration. Our method achieves a nearly linear performance scaling on manycore clusters and outperforms spatial-domain parallelization on a diverse set of benchmarks.

CCS Concepts

•***Computing methodologies*** → ***Physical simulation;***

1. Introduction

Significant progress has been achieved in visual simulation of cloth over the past decades [GHF*07, ZY01, BFA02, BW98]. Numerous algorithms have been proposed that achieve high accuracy and robustness for various 3D graphics applications, though real-time simulation remains illusive for complex simulation scenarios. Given recent advances in manycore and cloud computing, paral-

lel computing has emerged as a possible alternative to achieve the desired runtime performance. In this paper, we propose a novel method for parallelizing cloth simulation. Unlike previous methods, our method divides the workload in *time* domain that minimizes the communication overhead, thereby achieving much better scalability and higher performance gain over previous methods.

The key challenge in time-domain parallelization is to obtain or approximate the simulation states before the time-consuming simulation begins. We use a two-level mesh representation to address this time-dependency issue. Observing that a coarse-level mesh can be simulated at a much higher speed, our method runs a lower-resolution simulation using coarser meshes to approximate the state at each time step. After an appropriate remeshing process, the higher-resolution simulations using finer meshes can be run in parallel. To further refine the simulation results, we propose a practical technique to smooth the state transition from the low-resolution to high-resolution simulations. To recover the lost states, we make use of the coarse-level mesh and run several 'static' simulation steps before the high-resolution simulation starts. Experiments in Sec. 6 show that this technique can reduce the visual artifacts between temporal partitions. In order to balance the workload of each processor, we further develop an adaptive partitioning algorithm, which takes into account the varying time consumption of each frame caused by different contact configurations. We make use of the time measurements of previous frames in both mesh resolutions and determine the partition point based on the current estimation of the total running time.

To sum up, the key contributions of this work include:

- A time-domain parallelization algorithm supporting *adaptive meshes* with minimal communication overhead (Sec. 3);
- Load estimation and load balancing techniques that maximize the overall performance acceleration (Sec. 4);
- A practical state transitioning algorithm between low- and high-resolution simulations to recover details and ensure the visual quality of the simulated sequences (Sec. 5).

On a given set of benchmarks, our method achieves an unprecedented level of scalability in distributed CPU systems when compared to [ZFV04, NKT15]. Its performance gain is also higher than the GPU parallelization [TWT*16], while our approach offers the additional flexibility for coupling with adaptively remeshed cloth simulators. We also verify that given sufficient amount of processors, our method can achieve an average performance as fast as the low-resolution simulation, while obtaining simulation results similar to ones using high-resolution meshes. This method can be widely adopted in applications, where runtime performance is much more critical than accuracy, such as rapid design prototyping.

2. Related Work

In this section, we survey recent works on cloth simulation, parallelization techniques, and other related acceleration techniques for physics-based simulation.

2.1. Cloth Simulation

Simulation of cloth and deformable bodies has been extensively studied for a wide range of applications in different areas, from computer graphics, CAD/CAM, robotics and automation, to textile engineering. Due to their ability to take large time steps, implicit or semi-implicit methods [GHF*07, VMTF09, Zel05, BWK03] have been widely adopted after the seminal work by Baraff and Witkin [BW98]. However, most of these works focus on the serial simulation improvement and their runtime performances can be slow. We use one of the state-of-the-art simulation algorithms, ARCSim [NSO12], as the cloth simulator in our prototype implementation, but our parallelization technique does not rely on any specific simulation algorithm.

2.2. Time Parallel Time Integration Method

The scientific computing community have thoroughly studied parallelization techniques solving partial differential equations [EM12, SRK*12, RSE*13]. We refer readers to this survey paper by Gander et al. [GG] for more details. Cloth simulation is similar to the general time-evolution equations. However, there is a gap for these works to be directly applicable. Cloth simulation has coupled other non-PDE factors, such as the collision response due to continuous contacts with the human body. The standard collision response within Physically-based Modeling literature is usually an "empirical" impulse applied mainly on the boundary cases, where the cloth is about to collide with the body or within a pre-defined 'threshold' neighborhood. Traditional solutions [EM12] use an arbitrary initial guess (e.g.$\mathbf{X}_t = \mathbf{X}_0$) for each of the time step and try to update the overall solution using a fixed point iteration. The discontinuity introduced by collision not only prevents the method from solving the fixed point problem in Newton's method (calculating derivatives of the conditional term determined by variables to be solved), but also prevents most of the collision response algorithm from obtaining stable and correct results (a severe interpenetration of $\mathbf{X}_t = \mathbf{X}_0$ at time t that can hardly be handled). This special characteristic of cloth simulation makes it challenging to apply methods solving pure integrations (where the solution space is often regular) such as PFASST [EM12], due to collision-induced discontinuities.

2.3. Parallel Cloth Simulation

Parallelization is a popular, practical way to achieve performance improvement. Several parallelization techniques for cloth simulation have been proposed. [WY16, FTP16] proposed GPU-based simulation methods for elastic bodies. [MRB*99, RRZ00, KB04, TB06, ZFV02] proposed different types of spatial parallelization but they all suffer from severe sub-linear scalability due to large communication overhead. [NKT15] improved the work from [AVGT12] using Asynchronous Contact Mechanics and reduced the communication by proposing a locality-aware task assignment, which first scaled more than 16 cores. [TWT*16] implemented a GPU-based simulation pipeline. Their method has achieved an impressive speedup of 58 times, which is comparable to the performance of our method on a 64-core cluster.

The main difference between other parallelization methods and ours is that we decompose the simulation task in *time* domain. Partitioning in time domain significantly reduces the communication cost in distributed systems, thereby offering a considerable speedup. To the best of our knowledge, our method is the *first time-domain parallelization* algorithm for cloth simulation that can be coupled with *adaptive remeshing schemes*.

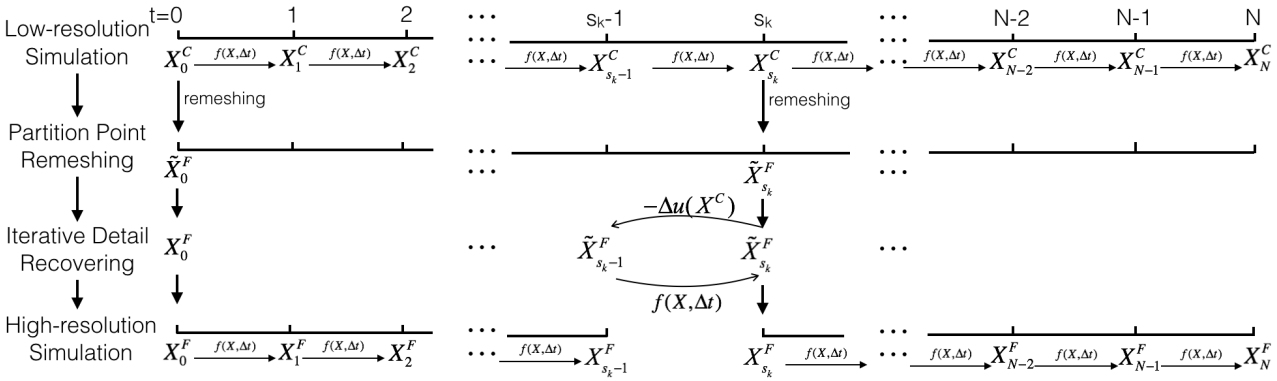


Figure 2: *An overview of our method. We first simulate the cloth mesh in low resolution, obtaining the approximated states* $\mathbf{X}^C_k$. *After we select the starting point in time for each processor* s_k *(Sec. 4), we use the upsampling function to generate the initial states* $\tilde{\mathbf{X}}^F_{s_k}$ *and recover the detail information iteratively (Sec. 5). Lastly, we simulate the entire sequence in parallel, given the starting states* $\mathbf{X}^F_{s_k}$.

2.4. Hierarchical Structures and Multi-level Methods

Multi-level algorithms have offered significant performance improvement on various simulation problems. Tamstorf et al. [TJM15] proposed a multi-grid method to speed up the cloth simulation. Bergou et al. [BMWG07] developed a tracking solver for rapid interaction in animation. They set up a two-level mesh representation and used the desired coarse level animation to guide the fine level one by applying constrained dynamics. Our method builds on top of their work to ensure the low-res consistency of the results. Recent works [MC10, WHRO10, RPC*10] generate high-resolution wrinkles from low-resolution cloth. Our method is a physically-aware approach; it's more diverse and realistic compared to those work. Ours is more of an intermediate trade-off between time-consuming simulation and physically-unaware wrinkle synthesis. We use a hierarchical mesh representation to approximate the states of the cloth mesh at each time step, before transitioning to computationally expensive high-resolution simulations on fine meshes.

2.5. Mesh Upsampling

Mesh upsampling algorithms are widely explored from geometrical approaches [SZD*98, DKT98, Loo87] to data-driven methods [KGBS11, FYK10]. Our method needs a specific mesh upsampling function to transfer the (approximated) state of the simulated cloth from low-resolution to high-resolution. While classic subdivision methods [Loo87] cannot generate high-resolution details, data-driven ones [KGBS11, FYK10] depend largely on the specific configuration in the training data, and as a result, can generate interpenetrations when applying to arbitrary scenarios. For generality, we do not assume any specific upsampling function. Instead, we introduce an iterative detail-recovering approach described in Sec. 5 in order to account for the lost details in the low-resolution mesh. In our experiment, we use an adaptive remeshing method in [NSO12] for its flexibility of use and a straightforward, linearly-interpolated subdivision for fast error computation.

3. Overview

In this section we give an overview of our approach. We define the problem formally before we introduce the basic idea of the method.

Problem Statement: Given the initial state of a cloth mesh, $\mathbf{X}_0$ (inclusive of both position and velocity), generate a sequence of cloth states $\mathcal{V} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ that characterize the cloth interaction with the given environment, using a time step Δt and a simulation function $\mathbf{X}_{k+1} = f(\mathbf{X}_k, \Delta t)$.

Fig. 2 shows the overall pipeline of our algorithm. The key idea of this method is to partition the time domain of the cloth simulation rather than the spatial domain of the simulated cloth. In order to obtain the (approximated) mesh state without full simulations, we propose a two-level hierarchy representation. We simulate the cloth mesh $\mathbf{X}^C$ at a coarser level with much lower computation and determine the partition point S (in time) according to the algorithm described in Sec. 4 before we simulate the entire high-resolution sequence $\mathbf{X}^F$ at the finer level in parallel.

The fine-level mesh at the starting point of each temporal partition is obtained by the corresponding coarse-level mesh using an upsampling/remeshing function $u(\mathbf{X}^C)$. However, the finer mesh may be quite different from the coarse one after remeshing because high frequency information $\mathbf{X}^D$ is not stored in the coarse-level mesh. Therefore, we design a practical state-transitioning technique to recover the lost details to the extent possible, before the high-resolution simulation begins. This state-transitioning method will be discussed in Sec. 5. We list the notations used in this paper in Table 1.

3.1. Two-Level Mesh Hierarchy Representation

Ideally we want to divide the whole simulation process into several temporal partitions so that we can simulate each partition in parallel and independently. However, since the mesh state at step k, $\mathbf{X}_k$, is determined by the state at previous step $\mathbf{X}_{k-1}$, we do not know the exact intermediate states until we finish the simulation from step 0 to step k. Here we use the hierarchical mesh representation to address this time-dependency problem. We maintain two sets of simulated meshes, $\mathbf{X}^C$ and $\mathbf{X}^F$, which represent the low- and high-res(olution) simulation states using the coarse- and fine-level meshes, respectively. We can recover the high-res state from the low-res one by a user-defined upsampling function: $\tilde{\mathbf{X}}^F = u(\mathbf{X}^C)$.

Note that the obtained high-res state from the fine mesh, $\tilde{\mathbf{X}}^F$, is only an approximation of the exact state $\mathbf{X}^F$. But, for simplicity,

Table 1: ***Notations and definition of our method.***

NOTATION	DEFINITION
$\mathbf{X}_k$	state of the cloth at step k
$\mathcal{V}$	output sequence of states
N	simulation sequence length
Δt	specified time step
$f(\mathbf{X}_k, \Delta t)$	one-step simulation
$f^i(\mathbf{X}_k, \Delta t)$	i-step simulation
$\mathbf{X}^C$	coarse level state
$\mathbf{X}^F$	exact fine level state
$\mathbf{X}^D$	state difference between the two level states
$\tilde{\mathbf{X}}^F$	approximated fine level state
$u(\mathbf{X}^C)$	upsampling function
p	number of processors
S	ordered set of starting points for parallelization
s_j	starting point of the jth processor
K	coarse-to-fine ratio

we assume that $\mathbf{X}^F = \tilde{\mathbf{X}}^F$ in this section. Further state refinement is discussed in Sec. 5.

Due to the fact that the simulation using a coarse mesh is significantly faster than the one using a fine mesh, we can obtain low-res states $\{\mathbf{X}^C_1, \ldots, \mathbf{X}^C_N\}$ in a relatively small amount of time. We further choose p starting points $S = \{s_0 = 0, s_1, \ldots, s_{p-1}\}$ in time for p processors, according to our partitioning algorithm to be discussed in Sec. 4.1, and run the high-res simulation using the fine mesh in parallel:

$$\mathbf{X}^F_k = \begin{cases} \tilde{\mathbf{X}}^F_k & k \in S \\ f^{k-s_j}(\mathbf{X}^F_{s_j}, \Delta t) & s_j < k < s_{j+1} \end{cases} \tag{1}$$

where

$$f^i(\mathbf{X}_k, \Delta t) = \begin{cases} f(f^{i-1}(\mathbf{X}_k, \Delta t), \Delta t) & i > 1 \\ f(\mathbf{X}_k, \Delta t) & i = 1 \end{cases} \tag{2}$$

for running i steps of simulation.

4. Time Domain Parallelization

In this section we will describe our parallelization technique. We solve the partitioning problem from the simplest case to the most complex one, in order to balance the workload of each processor.

4.1. Static Temporal Partitioning

A straightforward approach for the partition problem is to divide the time domain into p temporal segments of the same length:

$$s_j = \lfloor \frac{N}{p} j \rfloor \tag{3}$$

Assuming that every simulation step using the fine mesh takes the same amount of time, the overhead of this partition schedule is the time spent in simulation using the coarse mesh. To further simplify the case, we take another assumption that the simulation speed at the low-res level is K times as fast as high-res level. We can estimate the speedup as:

$$\eta_1 = \frac{KN}{K\frac{N}{p} + (p-1)\frac{N}{p}} = \frac{Kp}{K+p-1} \tag{4}$$

Note that in the low-res simulation using a coarse mesh there is no need to continue the simulation after we reach s_{p-1}. Therefore, the time spent on low-res simulation is $(p-1)\frac{N}{p}$.

One improvement of the straightforward approach is that we can start the high-res simulation in parallel, as long as the corresponding starting point is ready. Intuitively, we want all processors of the system to finish their jobs at the same time to achieve a good workload balance and the best speedup possible. This objective can be attained by adjusting the starting points so that the processor which starts earlier takes a longer part to simulate. Taking the same assumption, we arrive at a load-balancing equation:

$$s_j + K(s_{j+1} - s_j) = s_{j+1} + K(s_{j+2} - s_{j+1})$$

Recall that K is the ratio between the high- to low-res simulation time, s_j, s_{j+1}, and s_{j+2} are the starting point for simulation on the processors $j, j+1$, and $j+2$, respectively. This equation yields:

$$s_j = \lfloor \frac{1-q^j}{1-q^p} N \rfloor \tag{5}$$

where $q = 1 - \frac{1}{K}$. The speedup can then be expressed as:

$$\eta_2 = \frac{KN}{K(s_1 - s_0)} = K - K(1-1/K)^p \approx p - \binom{p}{2}\frac{1}{K} \tag{6}$$

This is a tighter bound than Eqn. 4, as p approaches to K. The key reason behind the sub-linear speedup is that the overhead ratio to the original computation is $1/K$. In practice, the ratio between high- to low-res simulation time can be controlled by the user and can usually reach 100∼200 using the method described in Sec. 4.3, which is sufficient for running on a large distributed system.

4.2. Adaptive Partitioning

In the discussion above, we consider K as a known constant throughout the entire simulation process. However, it is highly unlikely that this would be the case. First of all, remeshing in the simulation run leads to a varying number of vertices and thus a dynamically changing size of the linear system. Secondly, the computational cost can vary considerably, even with the same mesh size, due to collision queries. Recent studies [TWT*16] show that collision detection and response can take up to 80% of the total running time. Moreover, the difference of per-step runtime is also dominated by the collision response and the size of the adaptive mesh, which are largely related to the object granularity. It has much more impact in the high-resolution than the low-resolution, which K accounts for as well. Therefore, the ratio of high- to low-res simulation time varies and the exact number is usually unknown.

A fixed partitioning scheme can become unstable and sensitive to these variations, resulting in load imbalance. One common solution is to cut down the jobs into more smaller tasks so that the imbalance can be reduced by dynamic job scheduling scheme. This method surely works, but it will have large extra overhead due to job scheduling and required preprocessing time (Sec. 5), and extra hand-tuned granularity parameter to optimize the performance. Since we want to avoid any unnecessary computational overhead, we here propose an adaptive partitioning algorithm.

Suppose that we have simulated up to step n using the coarse

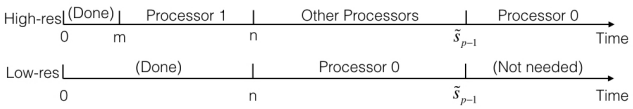


Figure 3: *Adaptive partitioning Algorithm. We estimate the ratio of high-to-low-res simulation time, $\tilde{K}$, according to the runtime data we observe so far ([0,m] in High-res on Processor 1 and [0,n] in Low-res on Processor 0). The objective is to predict the future running time (marked by 'Processor 0' and 'Processor 1' respectively) to be as close as possible to the actual time.*

mesh, when the first high-res parallel simulation with the same starting time has completed m steps, where $m < n$. Let $T_C(m)$ and $T_F(m)$ denote the running time of the previous m steps using the coarse and fine meshes, respectively. Then, the ratio of the high-to-low-res simulation time, $\tilde{K}$, can be approximated as:

$$\tilde{K} = \frac{T_F(m)}{T_C(m)} = \frac{T_C(n)}{T_C(m)} \tag{7}$$

Since these numbers may vary, it is not appropriate to determine the global partition points using current approximations. Instead, we use them to determine if we should perform a cut on step n, i.e. whether n should be s_1 or not. Fig. 3 gives a visualization of the process. The objective of the partitioning algorithm is that the total running time on the processor 0, which performs the low-res simulation and the last part of the high-res simulation, is equal to the running time of the current parallel simulation that performs the high-res simulation using a fine mesh from step 0 to step n. This relation can be formulated as:

$$T_C(\tilde{s}_{p-1}) + (T_F(N) - T_F(\tilde{s}_{p-1})) = T_F(n) \tag{8}$$

where $\tilde{s}_{p-1}$ is the estimated starting point of the last partitioned segment. We use the method described in Sec. 4.1 to obtain this parameter. We further approximate Eqn. 8 to:

$$n = \frac{N}{\tilde{K}} + \frac{\tilde{K}-1}{\tilde{K}}(N - \tilde{s}_{p-1}) \tag{9}$$

by assuming stable parameters in the remaining simulation:

$$T_F(j) = \tilde{K}T_C(j) = \tilde{K}T_C(1)j \quad \text{for any j} \tag{10}$$

Since n is increasing while $\tilde{K}$ and $\tilde{s}_{p-1}$ can be considered stable compared to n, Eqn. 9 can be defined at some point in $1 \leq n \leq N - p$. The remaining cut can be completed recursively. Algorithm 1 shows the pseudocode of this method. $\tilde{K}$ and $\tilde{s}_{p-1}$ here are approximated values used only for this cut. They can vary during the simulation, which will guide our partition algorithm to have adaptive cuts, instead of fixed ones in Sec. 4.1.

In practice, the overall performance using adaptive partitioning is similar to that using static partitioning when the user can manually select the best K value for the simulation scenario. This algorithm generally offers the advantage of dynamically estimating the ratio of the high-to-low-res simulation time, so the user does not need to hand-tune this parameter for the best possible speedup.

4.3. Analysis on Performance Scalability

As discussed in the previous sections, the scalability of this time-domain partitioning method for parallel cloth simulation depends

Algorithm 1 - Adaptive Partitioning

Require: N, p, X_0^C
1: $n \leftarrow 0$
2: start fine level simulation from step 0 on Processor 1
3: **while** true **do**
4: $n \leftarrow n+1$
5: obtain X_n^C from X_{n-1}^C
6: $m \leftarrow$ steps finished by Processor 1
7: calculate $\tilde{K}$, $\tilde{s}_{p-1}$ from Eqn. 5 and 7
8: **if** condition of Eqn. 9 is met **then** break
9: $t_1 \leftarrow n$
10: Control Processor 1 to stop at Step n
11: Recursively partition remaining N-n steps with p-1 processors

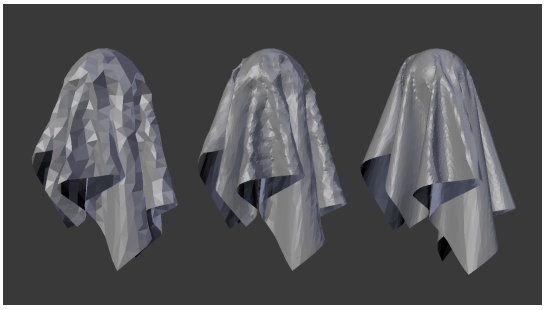


Figure 4: *An example of the coarse mesh $\mathbf{X}^C$, intermediate mesh, $\tilde{\mathbf{X}}^F$, and the fine mesh, $\mathbf{X}^F$, after iterative corrections.*

largely on the general runtime ratio between the high- to low-res simulation time, K. Since we perform a low-res simulation using a coarse mesh and a parallel one using a fine mesh, the low-res running time is a computational overhead for all processors and thus the speedup before any improvement is $\frac{K}{1+K/p} = \frac{Kp}{K+p}$. The ideal case of perfect workload balance, η_2, is discussed in Sec. 4.1, hence the actual performance of Algorithm 1 in a specific scenario, η_3, has the following theoretical bound:

$$\frac{Kp}{K+p} < \eta_1 \leq \eta_3 \leq \eta_2 < K \tag{11}$$

Therefore, the higher the K value is, the higher the overall performance gain of our method would be. One common way to increase K is to control the number of total mesh triangles by limiting the smallest possible size of each triangle in the low-resolution level. The other way is to enlarge the time step of the low-res simulation, since it is the common overhead of all processors and should aim for faster speed rather than smaller discretization errors. A properly chosen large time step can improve the overall performance with minimal impact on the simulation results. With the coarsening techniques in space and time domains, K can be sufficiently large to obtain good scalability in large distributed systems.

5. Smooth State Transitioning

As mentioned in Sec. 4, the high-res simulation state approximation $\tilde{\mathbf{X}}^F = u(\mathbf{X}^C)$ is not the same as the exact state $\mathbf{X}^F$ using the fine mesh, the reason of which is that the high frequency information needed to reconstruct the states of the fine mesh is missing in the estimated states of the simulation using the coarse mesh. Therefore,

if we take $\tilde{\mathbf{X}}^F$ directly as the starting state of the parallelized simulation, error $\mathbf{e} = E(\tilde{\mathbf{X}}^F, \mathbf{X}^F)$ will occur, since the high-frequency information is lost. Although $\mathbf{e}$ will vanish as the detail of the mesh is recovered by the simulation, another error will appear at the beginning of the subsequent partition after the end of the current one. (Here we focus on the actual visual effect instead of the L2 distance of each vertex. The error of our specific goal can be defined as the smoothness of the cloth.) Thus, this error will appear as a 'popping visual artifact' in the final concatenated sequence of the cloth simulation. Fig. 4 shows an example of the inaccurate starting mesh (middle) obtained from the corresponding coarse level mesh (left), which causes a popping visual artifact because the error compared to the actual state (right) is large enough to be visible.

One straight-forward method is to apply global smoothing optimization as a post-processing step. However, this space-time optimization is too time consuming to be used in speed demanding applications. As mentioned before, Bergou et al. [BMWG07] used constrained dynamics for fine level simulation to match with the coarse level motion. We employ this method to prevent the high-res simulation from diverging too far from the low-res one. However, the high-frequency detail information would be still missing at the transition point. Inspired from the observation that the visual error will be eliminated during the simulation, we propose an iterative refinement technique that can recover as much as possible the high-frequency detail of the cloth from the low-res simulation using the coarse mesh.

5.1. Iterative Detail Recovery

Consider the mesh state at the consecutive step points $\mathbf{X}^C_{k-1}$ and $\mathbf{X}^C_k$. The fine-level mesh can be regarded as the sum of the low-frequency coarse mesh and the high-frequency detail:

$$\mathbf{X}^F = u(\mathbf{X}^C) + \mathbf{X}^D \tag{12}$$

Assuming that the time step is sufficiently small and the detail does not change much between two simulation steps, we have:

$$\mathbf{X}^F_{k-1} - u(\mathbf{X}^C_{k-1}) \approx \mathbf{X}^F_k - u(\mathbf{X}^C_k) = \mathbf{X}^D_k \tag{13}$$

The idea here is to approximate $\mathbf{X}^F_{k-1}$ using $\mathbf{X}^C_{k-1}$, $\mathbf{X}^C_k$ and $\tilde{\mathbf{X}}^F_k$. From Eqn. 13 we have:

$$\tilde{\mathbf{X}}^F_k = f(\tilde{\mathbf{X}}^F_{k-1}, \Delta t) \tag{14}$$

$$\approx f(\tilde{\mathbf{X}}^F_k - u(\mathbf{X}^C_k) + u(\mathbf{X}^C_{k-1}), \Delta t) \tag{15}$$

Note that Eqn. 15 can be considered as an updated version of Eqn. 14. By subtracting the upsampled change of the state as a backward step and the simulation itself as a forward one, we can compute $\tilde{\mathbf{X}}^F_k$ iteratively. Algorithm 2 below shows the iterative detail recovery process. We run this algorithm at each of the transition point as a pre-processing step before the high-res simulation begins.

Algorithm 2 - Iterative Detail Recovery

Require: $\mathbf{X}^C_{k-1}$, $\mathbf{X}^C_k$ $(k \in S)$

1: $\tilde{\mathbf{X}}^F_k \leftarrow u(\mathbf{X}^C_k)$
2: **while** not reaching maximum iteration **do**
3: $\quad \tilde{\mathbf{X}}^F_{k-1} \leftarrow \tilde{\mathbf{X}}^F_k - u(\mathbf{X}^C_k) + u(\mathbf{X}^C_{k-1})$
4: $\quad \tilde{\mathbf{X}}^F_k \leftarrow f(\tilde{\mathbf{X}}^F_{k-1}, \Delta t)$ with constraints introduced by TRACKS [BMWG07]
5: $\mathbf{X}^F_k \leftarrow \tilde{\mathbf{X}}^F_k$

5.2. Convergence and Continuity

Taking the advantage of the constraint-based tracking solver introduced by Bergou et al. [BMWG07], this iterative algorithm can be proved to have convergence guarantee. We show the proof in Appendix A. It is not guaranteed that the convergence point is exactly the same as the high-res simulation result. However, due to the enforcement of the tracking constraint, the difference compared to the result at the previous step will be $O(\Delta t)$, which means that there will be very little discontinuity and in most practical cases they are invisible. We show several results in the supplementary video.

5.3. Iteration Number Estimation

The number of iterations needed for convergence, according to the proof, is largely related to the strength of the coarse-level constraint (in other words, the coarse-to-fine ratio K), since it provides the damping force to the system. Additionally, given a fixed upsampling scale (K), the iteration number is also related to a) the stiffness and density of the cloth, and b) the time step Δt. We use an qualitative estimation in Appendix B and directly gives out our approximation result here as $c_0\sqrt{m_s/\xi}/\Delta t$, where m_s is the density and ξ is the Frobenius norm of the stretching and bending stiffness matrix in [WOR11]. We use $c_0 = 10$ across all of our experiments. In practice, the iteration can also end when no large difference is detected between current and previous results. We found that using our estimation number the difference threshold can be as small as 10^{-3} relative to the scale of the cloth.

In each of the temporal partition, we add an extra simulation steps of $c_0\sqrt{m_s/\xi}/\Delta t$ to refine the starting state, so the total ideal performance gain due to parallelization is

$$\eta = \frac{N}{c_0\sqrt{m_s/\xi}/\Delta t + N/\eta_2} \tag{16}$$

Given a cloth material configuration with fixed m_s and ξ, η will have an upper-bound of η_2 if $c_0\sqrt{m_s/\xi}/\Delta t \ll KN/\eta_2$. This can be easily satisfied since the duration $N\Delta t$ is usually from a few seconds to many minutes, and $c_0\sqrt{m_s/\xi}$ is usually smaller than 1.

5.4. Implementation Details

There are some minor details in the implementation of the approach. When we take a larger step in the low-resolution simulation, we estimate the change of the state in the corresponding high-res step $u(\mathbf{X}^C_k) - u(\mathbf{X}^C_{k-1})$ by linearly interpolating the states in between. The same method is also used in the adaptive partitioning method described in Sec. 4.2. The recovery iterations also count into the estimation of the current $\tilde{K}$, but do not count into the total number of steps, N, since there is no corresponding step in the low-res level and each processor has the same number of extra simulation steps, so the system still remains balanced. We regard $\tilde{K}$ as $+\infty$ if the first step of the high-res simulation is not finished at the time we determine n in Sec. 4.2. Note that the state $\mathbf{X}$ includes both the position and velocity components. We also refine the velocities in the upsampling phase. When using adaptive remeshing,

we obtain the new velocity as the average of the two vertices during edge splitting, following ArcSim [NSO12]. The change of the state is also computed correspondingly.

5.5. State Inconsistency

In the extreme cases where the high-resolution mesh is much finer than the low-res one, e.g. 1M versus 100, the shape of the cloth in that case is largely determined by the aggregated effect from details not captured by low-res simulation. Therefore, we cannot recover the exact detail as in the serially simulated one at the transition point, which is referred to as the 'state inconsistency problem'. Enforcing the high-res mesh to match the low-res one using the tracking solver [BMWG07] can effectively avoid this problem. So, it can lead the simulation result to follow the movement of low-res one instead, which limits this approach from accuracy-demanding usage in those extreme cases. However, for other usage such as rapid design prototyping, where environmental constraints are mild and K is reasonable, motion difference between two levels is small and we can indeed achieve visually plausible results with high speedup, which are shown in Fig. 10 and 11. Alternative methods to improve the speedup without harming the accuracy is also discussed later in Sec.6.5.

6. Results

Our method is tested on a large computing cluster with 526 compute nodes, each with 12-core (dual socket), 2.93 GHz Intel processors, 12M L3 cache (Model X5670), and 48 GB memory at 2:1 ratio IB interconnect, MPI for communication. We run one process in each of the cores (compactly assigned). We use up to 128 cores of this cluster to show the linear scalability provided by our theory and up to 512 cores to show the maximum possible speedup in large distributed systems. We could not test on a larger number of cores due to a core limit of 512 per job locally. We use the upsampling function by [NSO12] throughout all of our experiments except in Table 4, which uses linearly-interpolated subdivision for fast error computation. As stated in Sec. 5.5, this method cannot guarantee the same accuracy as full simulation, which often cannot guarantee the same accuracy as the physical systems. The objective of this work is to generate visually plausible simulation to provide rapid visual feedback for interactive applications, such as rapid design prototyping.

6.1. Parameter and Scenario Setting

As mentioned in Sec. 4.3, we control the general coarse-to-fine ratio by limiting the smallest mesh size and enlarging the time step of the low-res simulation. Specifically in all of our test cases, the smallest length size of the triangle in the low-res simulation is about 5 times as large as that in the high-res one. The number of iterations in each of the smoothing processes is set to be the same as that in Sec. 5.3. We use ARCSim [NSO12] as our base simulator, since it naturally supports adaptive mesh refinement with an efficient remeshing algorithm. Our method can be used in other CPU-based simulators using uniform meshes as well, as long as the upsampling algorithm is specified or implemented. All listed K in the following tables are averaged values across the entire simulation. We show scaling results using figures for clarity. Please refer to Appendix C for detailed data.

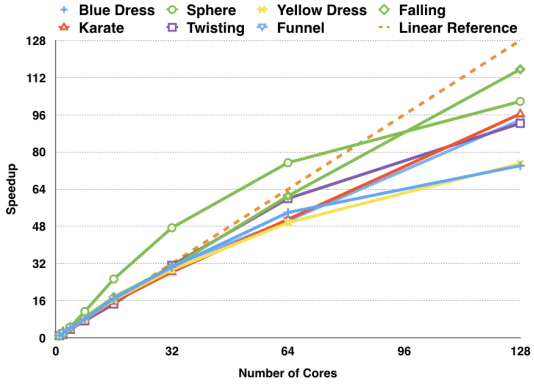


Figure 5: *Performance scaling result with large low-res time step. A nearly linear scalability is achieved.*

Table 2: *Results on a higher-resolution mesh. We run our system on meshes of higher resolution. Values in the table are the running time in minutes, while the numbers in the bracket are the corresponding speedup.*

Scenario	Blue Dress	Yellow Dress	Sphere	Falling	Karate	Twisting	Funnel
20K-94K	12.8(74.1)	26.8(75.0)	7.15(102)	61.5(116)	52(96.4)	28.8(92.3)	452(93.8)
80K-376K	74(99.6)	193(109)	30.2(178)	609(119)	599(103)	192(101)	942(108)

We use 7 different benchmarks to test the performance and the animation quality of our method: **Blue Dress and Yellow Dress** (Fig. 11(a,b)), **Sphere**(Fig. 11(c)), **Falling**(Fig. 11(d)), **Karate**(Fig. 1), **Twisting**(Fig. 10(a)) and **Funnel**(Fig. 10(b)). The default setting is 20 second simulation at the low-resolution time step of 0.02 sec using 128 cores. We extend the duration to 80 seconds and decrease the time step to make comparisons and validate our theoretical analysis on performance gain. Below are descriptions of each benchmark data.

To the best of our knowledge, previous works did not provide any code or experimental data to public, so the best known practice is to use the reported 'speedup data' in other works with similar scenarios, to minimize the difference due to computing platforms or implementation. We use the timing data of 'Two Cloths Draped' scenario from [NKT15] since it has similar settings as ours (cloth-object interaction), similarly with other benchmarks.

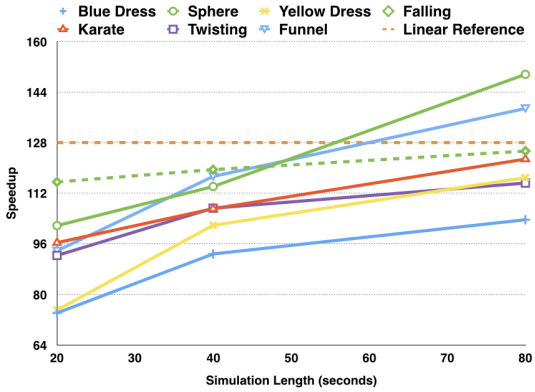


Figure 6: *Results with Increasing Length of the Simulation. A larger speedup is observed with longer duration of simulation.*

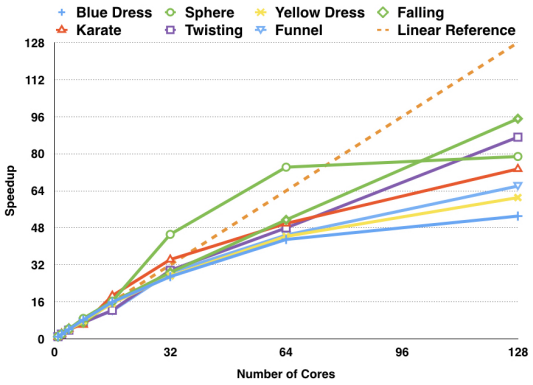


Figure 7: *Performance scaling result with small low-res time steps. Compared to Fig. 5, the speedup for cases with core number larger than 32 is decreased, due to the smaller time steps for low-res simulation.*

Table 3: *Comparison between different partition schemes. Values in the table are simulation runtime in seconds.*

Cores	8	16	32	64	128
Uniform partition runtime(s)	5533	3010	1042	684	631
Adaptive partition runtime(s)	4721	2568	928	565	532
Speedup (%)	117	117	112	121	119

6.2. Performance

Nearly linear scalability w.r.t. the number of cores. As indicated in Fig. 5, our method achieves a good scalability with an increasing number of processors. The reason of the super-linear speedup in the 'Sphere' scene is that it contains rapidly changing contacts with obstacles. When the cloth is free from contact after the sphere passes through, the remeshing algorithm of ARCSim failed to simplify the mesh effectively, spending an unnecessarily large amount of time simulating simple flat cloth. However, due to the nature of our two-level structure, we maintain a reasonably small number of mesh elements while preserving the quality, and therefore outperform the serial approach significantly. We tested our method on a higher-resolution mesh and observed an even better speedup (Table 2) due to the same reason.

Improved scalability with increasing simulation duration. We show in Fig 6 that the scalability of our parallelized cloth simulation improves as the duration of the simulation increases. Although the averaging effect of the remaining load imbalance may partially account for it, the most likely reason is from Eqn. 16. We have relatively small speedup in 128-core parallelization when simulating a 20-second simulation because the iterative detail recovery algorithm consumes a relatively large amount of time according to Eqn. 16. Since the overhead is not dependent on the duration of the simulation and our method is a time-domain parallelization technique, the performance gain improves as the length of the simulation increases due to a smaller portion of the overhead.

Performance impact on different choices of parameters. To verify our scalability analysis in Sec. 4.3 and 5.3, we further ran our benchmark with much smaller time steps in low-res simulation. As mentioned in Sec. 4.3, increasing low-resolution time step is one of the ways to increase the ratio of high-to-low-res simulation time, K. Fig. 7 shows that smaller time steps in low-res simulation leads to a sub-linear scaling in all datasets, starting from the 64-core configuration. Although the 'Sphere' dataset has a bigger K due to its simplicity, the scalability starts to degrade at 128 cores as well. The speedup still increases with the simulation duration (as shown in Table 7 of Appendix C). However, as it is more closely bounded by K, the gain factor is not as significant as that with large time steps. In practice, a large time step in low-resolution simulations is beneficial to the parallelization performance, but it is limited by (a) the embedded simulation method, (b) the duration of a single frame, and (c) the desired animation quality.

Performance impact on different partition schemes. Table 3 shows that by using our adaptive partitioning scheme, we achieve an average of about 120% speedup compared to the uniform partitioned one with the best chosen parameter. In cases such as rapid design prototyping, where the cloth is in continuous contact with obstacles, the parameter K remains relatively stable. However, it is still difficult to compute K before simulation begins, since it depends on the specific mesh and collision structure. Furthermore, it is best not to compute the parameter using the first few frames, since the cloth at the beginning can be under constrained without sufficient contact with the obstacles. Our adaptive partitioning method here serves as an on-the-fly parameter estimation algorithm in order to achieve good workload balance.

Low-res speed with high-res mesh on a large distributed system. We further test our method in extreme cases where K is relatively small compared to p, which is possible in practice when the computational resources are sufficient. The runtime result is shown in Table 4. Although we cannot achieve a speedup as high as 512 due to the limitation of K, we have actually met the upper bound. The serial low-resolution simulation has consumed most of the time so there is very little space to improve in our scheme.

Comparison with previous CPU parallelization work. We compare the performance of our method against other CPU parallelization techniques. Fig. 8 shows that in smaller-scale systems (less than 16 cores), our method can maintain a linear speedup with respect to the single-core system, scaling better compared to previous CPU-based methods using spatial-domain partitioning, e.g. 11x over 16 cores by [ZFV04]. For larger-scale systems (Fig. 9), we achieved about 50% more efficiency than previous methods such as [NKT15]. In these methods, the processors need to send the information to each other, typically several times, when solving the linear system, resulting in large communication overhead and limited scalability. In contrast, our method only needs to share the states from low-resolution simulations once. Therefore, our method can achieve greater scalability and efficiency in comparison.

In addition, we compare our method with the original embedded OpenMP version of ARCSim (Table 6 in Appendix C). Although a maximum of 2.69x is observed using OpenMP with 2 cores due to a better cache usage in the linear solver, the performance scaling is poor when adding more cores, which results from that the simulation algorithm does not parallelize the remeshing process due to memory access issues. Our method disables the OpenMP feature in the ARCSim. Since we parallelize the simulation in time domain, we can avoid memory access control problems, thereby achieving a better speedup.

Comparison with GPU-based parallelization. Using similar

Table 4: *Results in the extreme case. We use 512 cores to simulate these scenes. Values in the table are in seconds per frame. The error metric is relative curvature difference compared to serial results in percentage. We use linear interpolated subdivision for fast error comparison.*

Scenario	Blue Dress	Yellow Dress	Sphere	Falling	Karate	Twisting	Funnel
Time step(low-res)	1/200s				1/100s	1/50s	1/125s
Time step(high-res)	1/200s						1/500s
# of faces(low-res)	5K	6K	8K	6K	4K	4K	4K
# of faces(high-res)	80K	95K	131K	94K	58K	65K	65K
# of triangles(obstacle)	20K	20K	1280	15K	28K	762	4K
K	165	170	172	60	99	188	794
Low-res speed (serial 1-core)	0.6	0.79	1	1.2	0.83	0.22	0.32
High-res speed(OpenMP 12-core)	32.2	44.3	55.9	23.2	27.6	13.7	86.7
Our method	0.89	1.14	1.3	1.5	0.91	0.41	1.22
Error before detail recovery	11%	12%	3.2%	22%	29%	46%	16%
Error after detail recovery	4%	6%	0.6%	5%	9%	14%	7%

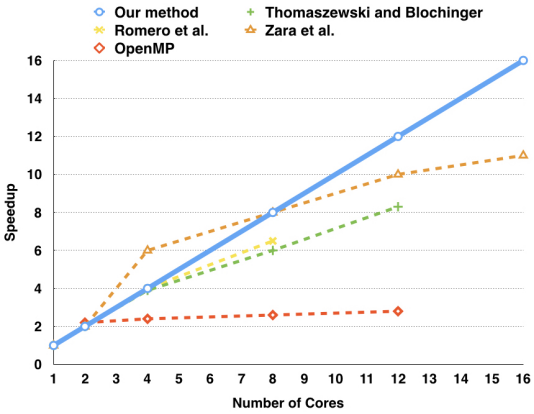


Figure 8: *Small scale parallelization comparison. Our method (in blue solid line) achieves a linear speedup, while others are limited by the communication overhead due to spatial domain partitioning.*

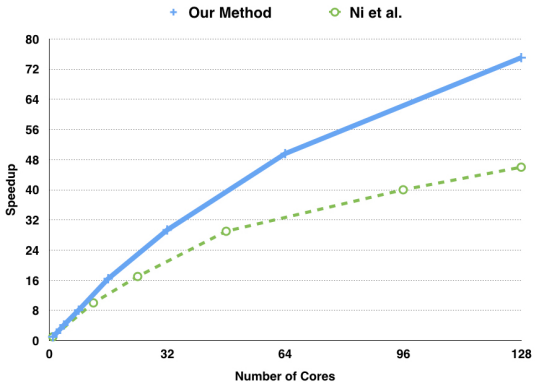
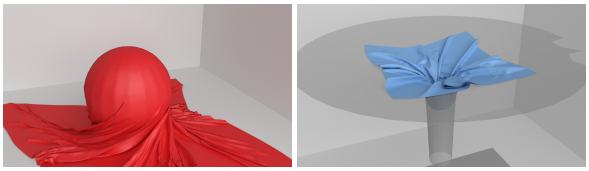


Figure 9: *Large scale parallelization comparison. Our method (in blue solid line) achieves about 50% higher efficiency than [NKT15] using dynamic workload balancing.*

Table 5: ***Comparison with GPU method [TWT*16].*** *Other than the scalable speedup gain with more cores, we are able to naturally support adaptive mesh during the simulation.*

Method	Speedup over sequential ArcSim [NSO12]	Supports Adaptive Mesh?
Tang et al. [TWT*16]	47-58x	No
Our method(64-core)	50-75x	Yes
Our method(128-core)	75-115x	Yes
Our method(512-core)	91-214x	Yes



(a) Twisting **(b)** Funnel

Figure 10: *More simulation results (best view with zoom-in in PDF).* We have achieved visually plausible and smooth results even in challenging cases involving frequent contacts.

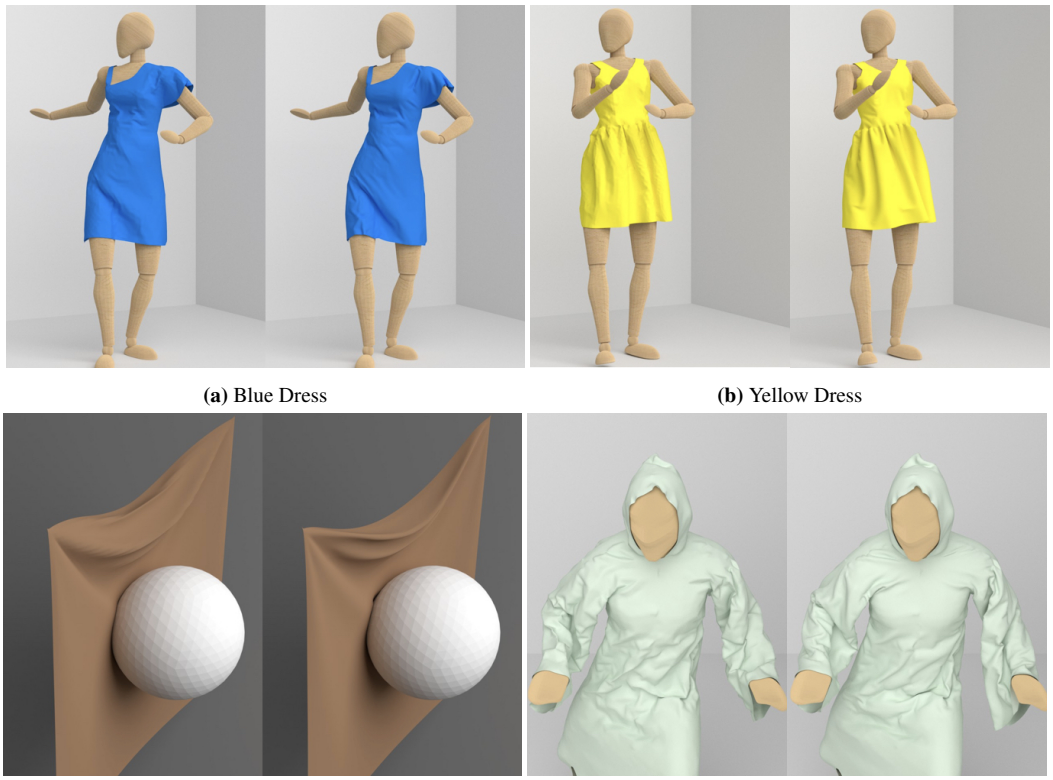
benchmarks as [TWT*16], the speedup of our method in a 64-core system configuration is up to 54x in practical scenarios compared to the original ArcSim implementation on a single-core system and achieves a performance gain comparable to the GPU parallelization of [TWT*16] (Table. 5). However, we have other distinctive strengths compared to the GPU method. We are the first work that can couple an adaptive mesh of varying dimensions during the simulation. We use the same number of triangles for performance comparison, but in practice we can produce similar visual granularity with much fewer triangles using adaptive mesh [NSO12], thereby making our method even faster. Moreover, our performance can be further improved using more cores and a longer simulation sequence, as shown in Fig 5 and 6.

6.3. Smoothness

Fig. 11 and Table 4 shows the results before and after the refining algorithm is applied. If directly using the results from the upsampling algorithm, the detail of the cloth is significantly different from the correct one and therefore introduces popping artifacts. After applying the iterative smoothing algorithm, the high frequency information is recovered. We use average curvature distance defined in Eqn. 17 to measure the error between the recovered mesh and the original, high-res one simulated using ArcSim on a single core.

$$E = \frac{\sum_{f_1,f_2 \in F} |curv(f_1,f_2) - curv(\tilde{f}_1,\tilde{f}_2)|}{\sum_{f_1,f_2 \in F} |curv(f_1,f_2)|} \tag{17}$$

where f_1, f_2 are two adjacent faces in the original mesh, and $\tilde{f}_1, \tilde{f}_2$ are two corresponding faces in our simulation result. We disable remeshing and use linearly-interpolated subdivision for fast comparison. A larger value of the curvature error indicates a sharper edge in the corresponding position and thus a potential artifact. Before our recovery method, a relative error up to 46% is observed, which can cause large 'popping' artifacts in the result animation



(a) Blue Dress **(b)** Yellow Dress

(c) Sphere **(d)** Falling

Figure 11: *Refining results (best view with zoom-in in PDF). The left image in each of the example is the upsampled mesh* ***without*** *detail recovery, which lacks high frequency details and causes 'popping' artifacts. The right one is the corresponding mesh* ***using our method****.*

(Fig. 11). By using our technique, the error has decreased by 2-5 times, which is a significant improvement. In the supplemental video, we show the improvements of the results in more detail, in which our method can achieve reasonably high-fidelity visual quality for parallelized cloth simulation.

6.4. Memory and Render Latency

The extra memory footprint introduced by our method is small compared to the high-res mesh. In our experiments, the low-res mesh storage is 5.5% of the high-res one. We do not render the low-res simulation in our method, and it actually starts at the same time with the first partition of the high-res one. Therefore, our method does not introduce any latency compared to the full-res simulation. In fact, we have achieved a 'pre-fetch' effect for the subsequent partitions due to the very fast, low-res simulation, thereby reducing any potential latency introduced by non-real-time simulation.

6.5. Limitations

There are some limitations with this method. First of all, the performance gain is bounded by the ratio of low- to high-resolution simulation time. Other than accelerating the simulation through parallelization in the temporal domain, we can additionally employ GPU implementation to further improve the overall gain. With a factor of 50x speedup from GPU [TWT*16] and a sufficient number of processors to parallelize the high-resolution simulation, it is possible to accelerate the performance even further. Secondly, the runtime of our method is bounded by a single-step high-resolution simulation time. This implies that at least one simulation step must take place in order to see the result. However, our method accelerates the overall performance, so we can actually achieve 'pseudo-interactivity', where the user can have a very fast visual feedback in parallel. Another possible direction is to implement a hybrid domain decomposition scheme, allocating some processors for spatial-domain parallelization to accelerate the single-step runtime. Our approach provides plausible visual results in practical real-time applications, like rapid design prototyping. However, as stated in Sec. 5.5, This approach may not be suitable in applications requiring high precision. In practice, the resulting cloth can sometimes appear slightly stiffer than the original one.

7. Conclusion and Future Work

In this paper, we introduce a novel temporal-domain parallelization method for practical cloth simulation such as rapid design prototyping. Taking the advantage of faster simulations on coarser meshes, we parallelize the cloth simulation in time with accelerated computation and minimal communication overhead. We also proposed an iterative detail recovery algorithm to minimize the visual artifacts due to the state transitioning from coarse to fine meshes. Our method outperforms existing CPU- and GPU-based parallelization techniques on a diverse set of benchmarks. It offers high efficiency and nearly linear scalability on large distributed systems, while maintaining high-fidelity visual simulation of the cloth. The scalability of our method is dependent on the ratio of low- to high-resolution simulation time, the length of the simulation, and per-

sistence of contacts with obstacles. Since this method utilizes only time-domain parallelization, a natural extension would be a hybrid decomposition scheme that may provide a potential usage in short-duration simulation or in circumstances with memory constraint.

Acknowledgement

This research is supported in part by NSF/CMMI Cybermanufacturing Program.

References

[AVGT12] AINSLEY S., VOUGA E., GRINSPUN E., TAMSTORF R.: Speculative parallel asynchronous contact mechanics. *ACM Trans. Graph. 31*, 6 (Nov. 2012), 151:1–151:8. URL: http://doi.acm.org/10.1145/2366145.2366170, doi:10.1145/2366145.2366170. 2

[BFA02] BRIDSON R., FEDKIW R., ANDERSON J.: Robust treatment of collisions, contact and friction for cloth animation. *ACM Transactions on Graphics (ToG) 21*, 3 (2002), 594–603. 1

[BMWG07] BERGOU M., MATHUR S., WARDETZKY M., GRINSPUN E.: TRACKS: Toward Directable Thin Shells. *ACM Transactions on Graphics (SIGGRAPH) 26*, 3 (jul 2007), 50:1–50:10. 3, 6, 7, 12

[BW98] BARAFF D., WITKIN A.: Large steps in cloth simulation. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques* (1998), ACM, pp. 43–54. 1, 2, 12

[BWK03] BARAFF D., WITKIN A., KASS M.: Untangling cloth. In *ACM Transactions on Graphics (TOG)* (2003), vol. 22, ACM, pp. 862–870. 2

[DKT98] DEROSE T., KASS M., TRUONG T.: Subdivision surfaces in character animation. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques* (1998), ACM, pp. 85–94. 3

[EM12] EMMETT M., MINION M. L.: Toward an Efficient Parallel in Time Method for Partial Differential Equations. *Communications in Applied Mathematics and Computational Science 7* (2012), 105–132. URL: http://dx.doi.org/10.2140/camcos.2012.7.105. 2

[FTP16] FRATARCANGELI M., TIBALDO V., PELLACINI F.: Vivace: A practical gauss-seidel method for stable soft body dynamics. *ACM Transactions on Graphics (TOG) 35*, 6 (2016), 214. 2

[FYK10] FENG W.-W., YU Y., KIM B.-U.: A deformation transformer for real-time cloth animation. In *ACM Transactions on Graphics (TOG)* (2010), vol. 29, ACM, p. 108. 3

[GG] GANDER M. J., GANDER M. J.: 50 years of time parallel time integration. 2

[GHF*07] GOLDENTHAL R., HARMON D., FATTAL R., BERCOVIER M., GRINSPUN E.: Efficient simulation of inextensible cloth. *ACM Transactions on Graphics (TOG) 26*, 3 (2007), 49. 1, 2

[KB04] KECKEISEN M., BLOCHINGER W.: Parallel implicit integration for cloth animations on distributed memory architectures. In *Proceedings of the 5th Eurographics conference on Parallel Graphics and Visualization* (2004), Eurographics Association, pp. 119–126. 2

[KGBS11] KAVAN L., GERSZEWSKI D., BARGTEIL A. W., SLOAN P.-P.: Physics-inspired upsampling for cloth simulation in games. In *ACM Transactions on Graphics (TOG)* (2011), vol. 30, ACM, p. 93. 3

[Loo87] LOOP C.: Smooth subdivision surfaces based on triangles. 3

[MC10] MÜLLER M., CHENTANEZ N.: Wrinkle meshes. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics symposium on computer animation* (2010), Eurographics Association, pp. 85–92. 3

[MRB*99] MAERTEN B., ROOSE D., BASERMANN A., FINGBERG J., LONSDALE G.: Drama: A library for parallel dynamic load balancing of finite element applicationsâŃŹ. In *European Conference on Parallel Processing* (1999), Springer, pp. 313–316. 2

[NKT15] NI X., KALE L. V., TAMSTORF R.: Scalable asynchronous contact mechanics using charm++. In *Parallel and Distributed Processing Symposium (IPDPS), 2015 IEEE International* (2015), IEEE, pp. 677–686. 1, 2, 7, 8, 9

[NSO12] NARAIN R., SAMII A., O'BRIEN J. F.: Adaptive anisotropic remeshing for cloth simulation. *ACM transactions on graphics (TOG) 31*, 6 (2012), 152. 2, 3, 7, 9

[RPC*10] ROHMER D., POPA T., CANI M.-P., HAHMANN S., SHEFFER A.: Animation wrinkling: augmenting coarse cloth simulations with realistic-looking wrinkles. In *ACM Transactions on Graphics (TOG)* (2010), vol. 29, ACM, p. 157. 3

[RRZ00] ROMERO S., ROMERO L. F., ZAPATA E. L.: Fast cloth simulation with parallel computers. In *European Conference on Parallel Processing* (2000), Springer, pp. 491–499. 2

[RSE*13] RUPRECHT D., SPECK R., EMMETT M., BOLTEN M., KRAUSE R.: Poster: Extreme-scale space-time parallelism. In *Proceedings of the 2013 Conference on High Performance Computing Networking, Storage and Analysis Companion* (2013), SC '13 Companion. URL: http://sc13.supercomputing.org/sites/default/files/PostersArchive/tech_posters/post148s2-file3.pdf. 2

[SRK*12] SPECK R., RUPRECHT D., KRAUSE R., EMMETT M., MINION M. L., WINKEL M., GIBBON P.: A massively space-time parallel N-body solver. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* (Los Alamitos, CA, USA, 2012), SC '12, IEEE Computer Society Press, pp. 92:1–92:11. URL: http://dx.doi.org/10.1109/SC.2012.6. 2

[SZD*98] SCHRÖDER P., ZORIN D., DEROSE T., FORSEY D., KOBBELT L., LOUNSBERY M., PETERS J.: Subdivision for modeling and animation. *ACM SIGGRAPH Course Notes 12*, 2 (1998), 43. 3

[TB06] THOMASZEWSKI B., BLOCHINGER W.: Parallel simulation of cloth on distributed memory architectures. In *Proceedings of the 6th Eurographics conference on Parallel Graphics and Visualization* (2006), Eurographics Association, pp. 35–42. 2

[TJM15] TAMSTORF R., JONES T., MCCORMICK S. F.: Smoothed aggregation multigrid for cloth simulation. *ACM Transactions on Graphics (TOG) 34*, 6 (2015), 245. 3

[TWT*16] TANG M., WANG H., TANG L., TONG R., MANOCHA D.: Cama: Contact-aware matrix assembly with unified collision handling for gpu-based cloth simulation. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 511–521. 1, 2, 4, 9, 10

[VMTF09] VOLINO P., MAGNENAT-THALMANN N., FAURE F.: A simple approach to nonlinear tensile stiffness for accurate cloth simulation. *ACM Transactions on Graphics 28*, 4 (2009), Article–No. 2

[WHRO10] WANG H., HECHT F., RAMAMOORTHI R., O'BRIEN J. F.: Example-based wrinkle synthesis for clothing animation. In *Acm Transactions on Graphics (TOG)* (2010), vol. 29, ACM, p. 107. 3

[WOR11] WANG H., O'BRIEN J. F., RAMAMOORTHI R.: Data-driven elastic models for cloth: modeling and measurement. In *ACM Transactions on Graphics (TOG)* (2011), vol. 30, ACM, p. 71. 6, 13

[WY16] WANG H., YANG Y.: Descent methods for elastic body simulation on the gpu. *ACM Transactions on Graphics (TOG) 35*, 6 (2016), 212. 2

[Zel05] ZELLER C.: Cloth simulation on the gpu. In *ACM SIGGRAPH 2005 Sketches* (2005), ACM, p. 39. 2

[ZFV02] ZARA F., FAURE F., VINCENT J.-M.: Physical cloth simulation on a pc cluster. In *4h Eurographics Workshop on Parallel Graphics and Visualization* (2002). 2

[ZFV04] ZARA F., FAURE F., VINCENT J.-M.: Parallel simulation of large dynamic system on a pc cluster: Application to cloth simulation. *International Journal of Computers and Applications 26*, 3 (2004), 1–8. 2, 8

[ZY01] ZHANG D., YUEN M. M.: Cloth simulation using multilevel meshes. *Computers & Graphics 25*, 3 (2001), 383–389. 1

Appendix A: Proof of Convergence of Algorithm 2

Theorem 1 Algorithm 2 can reach the convergence point when applying the coarse-level tracking constraints to the system, as long as $\frac{\partial \mathbf{F}}{\partial \mathbf{X}} = 0$ for external forces.

Proof We assume the whole system is running under the Forward Euler method:

$$\begin{pmatrix} \Delta \mathbf{x} \\ \Delta \mathbf{v} \end{pmatrix} = \Delta t \begin{pmatrix} \Delta \mathbf{v} \\ M^{-1}\mathbf{F}(\mathbf{X}) \end{pmatrix} \tag{18}$$

where $\mathbf{F}$ is the force function, and $\mathbf{X} = \begin{pmatrix} \mathbf{x} & \mathbf{v} \end{pmatrix}^T$ is the state of the cloth. Given the assumption that $\frac{\partial \mathbf{F}}{\partial \mathbf{X}} = 0$ for external forces, they have the same contributions for each iteration and are all canceled out by the subtraction $(\Delta u(\mathbf{X}_k^C))$ in Algorithm 2. So we only consider internal forces.

Since we only focus on one high-res simulation step here, we leave off the resolution superscript and replace the step number subscript by the iteration time. We denote the upsampled coarse-level difference by $\Delta \mathbf{X}_0 = \begin{pmatrix} \Delta \mathbf{x}_0 & \Delta \mathbf{v}_0 \end{pmatrix}^T$. Using the new notation, we have:

$$\begin{pmatrix} \mathbf{x}_i \\ \mathbf{v}_i \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{i-1} - \Delta \mathbf{x}_0 \\ \mathbf{v}_{i-1} - \Delta \mathbf{v}_0 \end{pmatrix} + \Delta t \begin{pmatrix} \mathbf{v}_{i-1} - \Delta \mathbf{v}_0 \\ M^{-1}\mathbf{F} \end{pmatrix} \tag{19}$$

We now regard the evolution from $\begin{pmatrix} \mathbf{x}_{i-1} & \mathbf{v}_{i-1} \end{pmatrix}^T$ to $\begin{pmatrix} \mathbf{x}_i & \mathbf{v}_i \end{pmatrix}^T$ as one full simulation step (instead of a backward-forward iteration), and only focus on the velocity equation (since the position can be derived from it):

$$\mathbf{v}_i = \mathbf{v}_{i-1} + \Delta t(M^{-1}\mathbf{F} - \Delta \mathbf{a}_0) \tag{20}$$

where $\Delta \mathbf{a}_0 = \Delta \mathbf{v}_0 / \Delta t$ is the corresponding acceleration value. Given that the internal forces are negative gradients of the potential energy, we have:

$$\begin{aligned} \frac{d^2\mathbf{x}}{dt^2} &= M^{-1}\mathbf{F} - M^{-1}M\Delta \mathbf{a}_0 && (21)\\ &= -M^{-1}\frac{\partial E}{\partial \mathbf{x}} - M^{-1}\frac{\partial M\Delta \mathbf{a}_0 \cdot \mathbf{x}}{\partial \mathbf{x}} && (22)\\ &= -M^{-1}\frac{\partial E}{\partial \mathbf{x}} - M^{-1}\frac{\partial E_0}{\partial \mathbf{x}} && (23)\\ &= -M^{-1}\frac{\partial \tilde{E}}{\partial \mathbf{x}} && (24) \end{aligned}$$

where we make up a form of potential energy (E_0) with constant gradients to unite the two components.

By computing the dot product with the velocity (of the previous iteration), we have:

$$\begin{aligned} \frac{d\mathbf{x}}{dt} \cdot M\frac{d^2\mathbf{x}}{dt^2} &= -\frac{d\mathbf{x}}{dt}\Big|_{(i-1)\Delta t} \cdot \frac{\partial \tilde{E}}{\partial \mathbf{x}}\Big|_{\mathbf{x}_{i-1}-\Delta \mathbf{x}_0} && (25)\\ &= -\frac{d\mathbf{x}}{dt} \cdot (\frac{\partial \tilde{E}}{\partial \mathbf{x}}\Big|_{\mathbf{x}_{i-1}} - \frac{\partial^2 \tilde{E}}{\partial \mathbf{x}^2}\Delta \mathbf{x}_0) && (26)\\ &= -(\frac{\partial \tilde{E}}{\partial t} - \frac{\partial^2 \tilde{E}}{\partial \mathbf{x}\partial t}\Delta \mathbf{x}_0) && (27)\\ &= -\frac{\partial}{\partial t}(\tilde{E} - \frac{\partial \tilde{E}}{\partial \mathbf{x}}\Delta \mathbf{x}_0) && (28)\\ &= -\frac{\partial \tilde{E}}{\partial t}\Big|_{\mathbf{x}_{i-1}-\Delta \mathbf{x}_0} && (29) \end{aligned}$$

or in a discrete form:

$$\mathbf{v}_{i-1} \cdot M\mathbf{a}_i = -\frac{\partial \tilde{E}}{\partial t}\Big|_{\mathbf{x}_{i-1}-\Delta \mathbf{x}_0} \tag{30}$$

This equation means that the whole system tends to decrease the sum of the potential energy: when $\tilde{E}$ is decreasing, the acceleration $\mathbf{a}_i$ will have roughly the same direction with the velocity; otherwise it will have the opposite one, makes the velocity direction turn around eventually. The coarse-level tracking constraint here serves as a damping component, which prevents the system from oscillation due to conservation of energy. It also prevents $\tilde{E}$ from infinitely decreasing since the coarse shape of the mesh is strictly preserved [BMWG07]. Therefore, after sufficient number of iterations the whole system will reach a balance where $\frac{\partial \tilde{E}}{\partial t} = 0$, and a stable result gives $\mathbf{v}_i = \mathbf{a}_i = 0$.

□

Note that although we have constraints on external forces, in most of the cases , they can be easily satisfied, such as gravitational forces and user-control impulse forces. Here we consider collision response as part of the constraint system, so it does not have impacts on the practical correctness. We use Forward Euler only for the simplicity of the expression in the proof. Actually we can derive the same form of Eqn. 20 using any other integrator (e.g. Backward Euler), during which the extra terms related to $\Delta \mathbf{v}_0$ (introduced by Backward Euler [BW98]) can be canceled out, eventually leaving $\Delta \mathbf{a}_0$. The main idea of the proof is that the system is conservative, regardless of the actual integrator, before adding extra damping constraints that ensures the final convergence. Upon convergence, the change in the high-res states (i.e. velocities and accelerations) will be the same as the change in the interpolated low-res states. This step, together with the position constraints by TRACKS, ensures the position and velocity difference between the high-res results at the boundary to be $O(\Delta t)$, smoothing out the visual popping artifact.

Appendix B: Iteration Number Estimation

We estimate our iteration number in a simplified 2-D spring-mass system. Suppose at $t = 0$ a string with length l is hanging horizontally, with both endpoints fixed. It is currently discretized as one single piece of 1-D string so the middle part of itself will not fall down. However, in the continuous real-world space, it is not in the equilibrium state and it has a residual energy of $O(l^2)$. This continuous case can actually be regarded as a string discretized to infinitely many small pieces. We define the residual energy as the difference of the potential energy between the current discretized one and the continuous one.

Subdividing the spring will bring the entire system closer to the actual continuous case (since the newly introduced vertices will fall down), so the residual energy will decrease. The spring system will start to bounce around upon discretization and we assume that there are damping forces in the system. After discretizing the spring into c pieces of equal length, the new system will have a residual energy of $O(l^2/c)$ when reaching the equilibrium state in the new discretization setting. If the system is in the critical damping condition, the energy will decrease by a factor of e after $t = \sqrt{m_s/\xi}$ seconds, where m_s is the mass of the spring and ξ is the stiffness. Therefore, the recovery time needed from the coarse level to the fine one is $O(\sqrt{m_s/\xi}\ln c)$.

In our case, we have $K = c^{O(1)}$ which depends on the embedded simulator and the collision state. Also we set $\ln K \leq 7$ to cover most of the cases. We use the density and the Frobenius norm of the stretching and bending stiffness matrix in [WOR11] to estimate $\sqrt{m_s/\xi}$. m_s typically ranges from 0.1 to 1, while the value of ξ is between 10 and 100.

Combining all of them above, we have an estimation of $c_0\sqrt{m_s/\xi}/\Delta t$ as the number of iteration steps needed.

Appendix C: Detailed Runtime Data

Table 6: *Performance Scaling Results.* Values in the table indicate the total running time of each setting in minutes, while the numbers in the bracket indicate the speedup with respect to the baseline performance on 1-core system. Compared to the embedded OpenMP implementation of ARCSim, we achieved much better performance and nearly-linear scalability.

Scenario	Blue Dress	Yellow Dress	Sphere	Falling	Karate	Twisting	Funnel
Time step (low-res)	1/50s						
Time step (high-res)	1/500s	1/500s	1/300s	1/200s	1/200s	1/200s	1/500s
# of triangles(low-res)	1375	1732	1576	6K	4K	4K	4K
# of triangles(high-res)	20K	40K	30K	94K	58K	65K	65K
# of triangles(obstacle)	20K	20K	1280	15K	28K	762	4K
K	368	381	564	187	158	212	1323
1-core(ARCSim)	947(1.00)	2010(1.00)	727(1.00)	7166(1.00)	5130(1.00)	2675(1.00)	56534(1.00)
2-core	467(2.03)	1003(2.00)	333(2.19)	3181(2.23)	1904(2.63)	1755(1.51)	26365(2.13)
4-core	221(4.29)	487(4.13)	155(4.70)	1545(4.60)	1427(3.51)	709(3.74)	13869(4.03)
8-core	118(8.06)	249(8.07)	64.1(11.4)	807(8.80)	624(8.03)	360.6(7.36)	6772(8.26)
16-core	56.1(16.9)	123(16.4)	28.7(25.3)	405(17.6)	321(15.6)	182(14.6)	3446(16.2)
32-core	31.1(30.4)	69(29.3)	15.4(47.4)	233(30.5)	177(28.3)	85.0(31.2)	1802(31.0)
64-core	17.6(54.0)	40.5(50.0)	9.65(75.4)	116(61.1)	98.2(51.0)	44.3(60.0)	1108(50.5)
128-core	12.8(74.1)	26.8(75.0)	7.15(102)	61.5(116)	50.4(99.5)	28.8(92.3)	596(93.8)
256-core	8.64(106)	15(134)	5.77(121)	31.3(123)	40.6(123)	26.1(102)	384(145)
512-core	8.76(105)	15.3(132)	5.98(117)	27.6(140)	34.0(147)	25.0(106)	167(333)
2-core(OpenMP)	512(1.85)	746(2.69)	336(2.17)	3947(1.80)	2357(2.13)	1574(1.69)	17845(2.37)
4-core(OpenMP)	474(2.00)	668(3.01)	317(2.29)	3656(1.94)	2110(2.37)	1064(2.49)	16569(2.56)
8-core(OpenMP)	447(2.11)	595(3.38)	315(2.30)	3014(2.36)	1871(2.68)	948(2.80)	14692(2.88)
12-core(OpenMP)	431(2.20)	585(3.43)	262(2.78)	1547(4.59)	1840(2.72)	913(2.91)	14448(2.93)

Table 7: *Results with Increasing Length of the Simulation.* All settings are the same as those in Table 6, except with varied simulation durations.

Scenario		Blue Dress	Yellow Dress	Sphere	Falling	Karate	Twisting	Funnel
Small	20 Seconds	17.3(53.1)	33.0(61.1)	8.87(78.8)	74.7(95.1)	68.2(73.4)	30.5(87.2)	642(66.0)
Time	40 Seconds	30.6(60.0)	53.4(75.4)	13.8(101)	142(100)	127(78.6)	59.7(88.8)	1224(69.2)
Steps	80 Seconds	57.4(63.9)	103(78.5)	28.4(98.6)	266(107)	245(81.8)	110(96.1)	2247(75.4)
Large	20 Seconds	12.8(74.1)	26.8(75.0)	7.15(102)	61.5(116)	52(96.4)	28.8(92.3)	452(93.8)
Time	40 Seconds	20.4(92.8)	39.5(102)	12.8(114)	119(119)	93.6(107)	38.1(107)	723(117)
Steps	80 Seconds	36.6(104)	68.8(117)	19.5(150)	227(125)	163(123)	54.2(115)	1221(139)

Table 8: *Performance Scaling Results with smaller time steps for low-res simulation.* All settings are the same as those in Table 6, but with the low-res simulation time steps decreased to twice as much as those in high-res simulation.

Scenario	Blue Dress	Yellow Dress	Sphere	Falling	Karate	Twisting	Funnel
Time step (low-res)	1/250s	1/250s	1/150s	1/100s	1/100s	1/100s	1/250s
Time step (high-res)	1/500s	1/500s	1/300s	1/200s	1/200s	1/200s	1/500s
K	116	137	267	117	87.9	113	257
1-core(ARCSim)	919(1.00)	2013(1.00)	699(1.00)	7101(1.00)	5012(1.00)	2654(1.00)	55939(1.00)
2-core	459(2.00)	1007(2.00)	333(2.10)	3188(2.23)	2128(2.35)	1280(2.07)	25256(2.21)
4-core	223(4.12)	517(3.89)	162(4.32)	1555(4.57)	1118(4.48)	693(3.83)	14679(3.81)
8-core	111(8.25)	258(7.81)	78.7(8.88)	1015(7.00)	814(6.16)	378(7.02)	7435(7.52)
16-core	57.7(15.9)	131(15.4)	42.8(16.3)	445(16.0)	268(18.7)	214(12.4)	4604(12.1)
32-core	34.1(27.0)	74.0(27.2)	15.5(45.2)	247(28.8)	146(34.3)	89.5(29.6)	1949(28.7)
64-core	21.3(43.0)	45.4(44.3)	9.42(74.2)	28.8(51.4)	100(49.9)	55.3(48.0)	1246(44.9)
128-core	17.3(53.1)	33.0(61.1)	8.87(78.8)	74.7(95.1)	52(96.4)	30.5(87.2)	846(66.0)
256-core	15.4(59.6)	29.3(68.6)	8.93(78.2)	43.3(89.1)	47.8(105)	26.4(100)	432(129)
512-core	14.8(62.1)	28.2(71.5)	9.08(76.9)	44.4(87.0)	40.5(124)	25.8(103)	240(232)