Improved scaling of temperature-accelerated dynamics using localization

Yunsic Shim and Jacques G. Amar
Department of Physics and Astronomy, University of Toledo, Toledo, Ohio 43606, USA

(Received 31 March 2016; accepted 17 June 2016; published online 6 July 2016)

While temperature-accelerated dynamics (TAD) is a powerful method for carrying out non-equilibrium simulations of systems over extended time scales, the computational cost of serial TAD increases approximately as N^3 where N is the number of atoms. In addition, although a parallel TAD method based on domain decomposition [Y. Shim *et al.*, Phys. Rev. B **76**, 205439 (2007)] has been shown to provide significantly improved scaling, the dynamics in such an approach is only approximate while the size of activated events is limited by the spatial decomposition size. Accordingly, it is of interest to develop methods to improve the scaling of serial TAD. As a first step in understanding the factors which determine the scaling behavior, we first present results for the overall scaling of serial TAD and its components, which were obtained from simulations of Ag/Ag(100) growth and Ag/Ag(100) annealing, and compare with theoretical predictions. We then discuss two methods based on localization which may be used to address two of the primary "bottlenecks" to the scaling of serial TAD with system size. By implementing both of these methods, we find that for intermediate system-sizes, the scaling is improved by almost a factor of $N^{1/2}$. Some additional possible methods to improve the scaling of TAD are also discussed. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4954996]

I. INTRODUCTION

Molecular dynamics (MD) is one of the most popular computational methods used to simulate the dynamics of materials at the atomistic level. However, due to the small time step for time evolution, MD is typically limited to times of the order of microseconds, while many crucial atomistic processes occur on much longer time scales. As a result, a variety of accelerated dynamics methods have been developed,[1–32] including hyperdynamics,[4,5,14–17,32] parallel replica dynamics,[6–9] temperature-accelerated dynamics (TAD),[10–13] metadynamics,[21–24] basin-searching,[26–31] and κ-dynamics.[25] In particular, TAD simulations of small systems have been shown to lead to speed-ups of 4–6 orders of magnitude compared to molecular dynamics. As a result, TAD has been used to study the long-time evolution in a variety of non-equilibrium systems in which activated events play a crucial role.[11,33–39] However, due in part to the fact that as in other accelerated dynamics methods there is typically no assumption of locality, TAD suffers from poor scaling with increasing system-size. Therefore, it is of interest to develop methods which can improve the scaling behavior.

In TAD, a basin-constrained MD simulation at a high-temperature $T_{\rm high}$ is used to accelerate the search for activated event transition pathways at a lower temperature $T_{\rm low}$. The corresponding activation energies are then used, along with the assumptions of harmonic transition-state theory[40] and a minimum prefactor $\nu_{\rm min}$ and confidence level $1-\delta$, to determine the first transition which will occur at the desired low temperature $T_{\rm low}$. To maximize the speed of the search for transition pathways, typically the ratio $\gamma = T_{\rm low}/T_{\rm high}$ is significantly less than 1.

As discussed in Ref. 12, in an ideal TAD simulation the ratio of the high-temperature MD time to the corresponding simulated low-temperature time is expected to scale (ignoring the effects of long-wavelength modes[41]) as $N^{1-\gamma}$, where N is the number of moving atoms. This implies that both the high-temperature MD computational work per unit of low-temperature time and the corresponding number of observed high-temperature transitions are expected to scale as $N^{2-\gamma}$. If the amount of computational work for each attempted high-temperature transition also scales as N, then this implies that the total time for a serial TAD simulation will scale approximately as N^{α} where $\alpha = 3-\gamma$.

In an effort to improve the scaling behavior, we have previously developed a parallel TAD (parTAD) method[42] based on domain decomposition along with our semi-rigorous synchronous sublattice (SL) algorithm.[43] Using this method we have demonstrated $\log(N)$ scaling with system-size. This method has also been used to carry out large-scale simulations of Cu/Cu(100) and Ag/Ag(100) growth[36] over extended time scales in order to explain the results of recent experiments.[44–46] A similar approach has recently been used to develop a synchronous sublattice parallel replica dynamics method for exascale computing.[47]

However, while parTAD simulations lead to excellent scaling with system-size, they suffer from two possible drawbacks. One drawback is that the size of activated events that can occur in the system is inherently limited by the spatial decomposition size rather than the total system size. In addition, increasing this size by a factor of f to increase the size of activated events leads to a further slowdown by a factor of $f^{3-\gamma}$. A second drawback is that, especially for small sublattice or spatial decomposition size, the cycle-time τ (corresponding to the time interval between local

communications between processors) needs to be carefully controlled to avoid errors in the dynamics. Therefore, it is of interest to develop additional methods to improve the scaling of serial TAD with system size.

In order to better understand the key factors which determine the scaling of TAD, here we first present the results of benchmark simulations for the scaling of serial TAD as well as for its different components, as a function of system-size for the case of Ag/Ag(100) growth and Ag/Ag(100) annealing. We note that due to the existence of large fluctuations, and depending on the system-size, averages have been taken over from 5-20 TAD runs. We then compare the scaling obtained in our simulations with theoretical predictions.[12] Based on these results, we have identified the two components of TAD which dominate the scaling behavior for medium-to-large system-sizes, e.g., the calculation of the activation barrier for each "new" high-temperature transition, and transition refining. By applying localization to both of these components of TAD, we find significantly improved scaling for localized TAD (LTAD). We also discuss the possibility of applying localization to other components of TAD in order to further improve the scaling.

The organization of this paper is as follows. We first review the TAD method and its expected scaling behavior in Sec. II. In Sec. III, we then present our benchmark results for the scaling of serial TAD obtained from simulations of Ag/Ag(100) growth and Ag/Ag(100) annealing. In Sec. IV, we first briefly describe our localized saddle-point (LSAD) method as well as our local refining (LREF) method. The possibility of on-the-fly transition detection is also discussed. We then present detailed scaling results obtained from LTAD simulations of Ag/Ag(100) growth and Ag/Ag(100) annealing and compare with our benchmark TAD scaling results. Finally, in Sec. V we summarize our results and discuss some additional possible methods to improve the scaling of TAD.

II. REVIEW OF TAD METHOD AND SCALING

We first briefly review the TAD method[10] as well as the expected scaling of each of its key components. As already noted, TAD uses a basin-constrained MD simulation at a high-temperature $T_{\rm high}$ to accelerate the search for transition pathways from a given initial state or energy basin and thus determine the first activated event which will occur at a lower temperature $T_{\rm low}$. In particular, assuming harmonic transition state theory,[40] one may show that for any high-temperature transition i observed after a high-temperature MD time $t_{i,\rm high}$, the corresponding extrapolated low-temperature event time $t_{i,\rm low}$ is given by

$$t_{i,\rm low} = t_{i,\rm high} e^{E_{a,i}(\beta_{\rm low}-\beta_{\rm high})}, \tag{1}$$

where $\beta = 1/k_{\rm B}T$, $k_{\rm B}$ is Boltzmann's constant, T is the temperature, and $E_{a,i}$ is the activation barrier for transition i, which is typically determined by a saddle-point search method, such as the climbing image nudged elastic band (cNEB) method.[48] After each observed high-temperature transition, the system is returned to the original basin and re-thermalized and decorrelated before continuing the high-temperature MD search for additional transition pathways. This process continues until the total high-temperature MD time $t_{\rm high}$ reaches the high-temperature stop-time $t_{\rm high,stop}$ given by

$$t_{\rm high,stop} = \frac{\ln(1/\delta)}{\nu_{\rm min}} \left(\frac{\nu_{\rm min} t_{\rm low,short}}{\ln(1/\delta)} \right)^{T_{\rm low}/T_{\rm high}}, \tag{2}$$

where $1-\delta$ is the confidence level (typically chosen to be greater than or equal to 90%), $\nu_{\rm min}$ is the assumed minimum prefactor for activated events, and $t_{\rm low,short}$ is the shortest extrapolated low-temperature event time found so far for the state. Once this time is reached, the transition corresponding to the $t_{\rm low,short}$ is accepted, the system is placed in the corresponding basin, and a new search for transition pathways is begun.

Figure 1 shows a flow chart of the serial TAD algorithm which describes the detailed steps and key elements which must occur before accepting an event at $t_{\rm low,short}$. TAD starts by thermalizing a minimized initial configuration or preparing another decorrelated MD configuration for a new transition search at temperature $T_{\rm high}$. If no transition is observed at the end of the corresponding "blackout" time (typically 1 ps) then the system is assumed to be thermalized and decorrelated, and a high-temperature MD simulation is carried out for a fixed number $n_{\rm MD}$ of MD steps (referred to as a *block*). Alternatively, if a transition is detected, then the system is

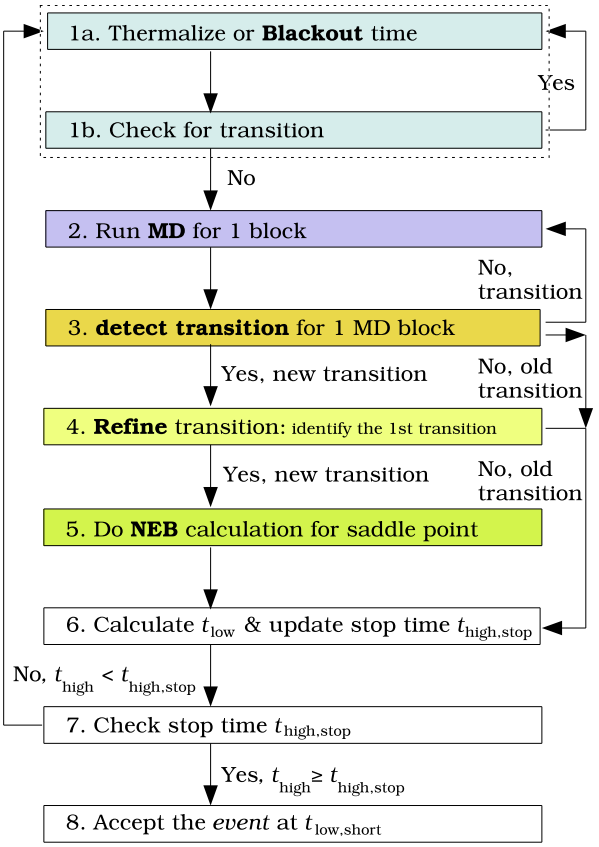

FIG. 1. Flow chart of serial TAD algorithm showing detailed steps required before accepting an event at $t_{\rm low,short}$. The dotted box represents the "blackout process" (see text) used to decorrelate the system after each attempted event. The five most time-consuming components (processes 1-5) are indicated by the use of bold text and different colors.

returned to the original basin and the *blackout* process is repeated. After each MD block, the final high-temperature MD configuration is minimized to determine if a transition to a new basin has occurred (*transition detection*). If the final minimized state is identical to the initial minimized state, then it is assumed that no transition has occurred, and another block of MD simulation is carried out and the total elapsed high-temperature MD time is increased by the block time. Alternatively, if a transition has occurred then the final state is compared with previously observed final states in order to determine if a new transition pathway has been found. If the minimized state corresponds to a possible new transition pathway, then this *transition detection* process is followed by *transition refining*, which involves a series of bisections and minimizations, in order to identify the first high-temperature transition which occurred during the given block (see an example in Fig. 6(a)). The last major step is then to carry out a *NEB* calculation for the corresponding transition in order to determine its activation barrier E_a, followed by calculating the corresponding extrapolated low-temperature time $t_{i,\mathrm{low}}$ and then checking and/or updating the high-temperature stop time. This whole process repeats until the event at $t_{\mathrm{low,short}}$ is accepted. Thus, the five major time-consuming parts in serial TAD are (1) blackout, (2) high-temperature MD, (3) transition detection, (4) transition refining, and (5) activation energy (NEB) calculation, as illustrated in Fig. 1.

As discussed in the Introduction, in an ideal TAD simulation in which all attempted events are immediately detected during the high-temperature MD, the ratio of the high-temperature MD time to the corresponding simulated low-temperature time is expected to scale[12] as $N^{1-\gamma}$. While this result was rigorously derived[12] using the probability distribution of basin escape times, it may also be obtained using a somewhat simpler argument. In particular, since the total event rate is proportional to N, we expect that $t_{i,\mathrm{high}}$ scales as $1/N$. Using Eq. (1) this implies that $t_{\mathrm{low,short}}$ also scales as $1/N$. Substituting in Eq. (2) then implies that $t_{\mathrm{high,stop}}$ scales as $N^{-\gamma}$. Dividing $t_{\mathrm{high,stop}}$ by $t_{\mathrm{low,short}}$ then implies that the ratio of the high-temperature MD time to the corresponding simulated low-temperature time scales as $N^{1-\gamma}$.

Using this result, the ideal theoretical scaling for each of the five main components of serial TAD may now be obtained. In particular, since the computational work per high-temperature MD step scales as N, we expect that the computational work for the high-temperature MD (step 2 in Fig. 1) should scale as $N^{2-\gamma}$. Since the number of high-temperature transitions per unit of high-temperature time is also expected to scale as N, this implies that the number of attempted transitions per unit of low-temperature time also scales as $N^{2-\gamma}$. Since the amount of computational work per attempted transition for the four remaining components of TAD, e.g., transition detection, transition refining, saddle-point calculation, and decorrelation/blackout also scales as N, then we expect that the computational work for each of these components will scale approximately as $N^{3-\gamma}$. As a result, the overall computational work for serial TAD is expected to scale as $N^{3-\gamma}$. Nevertheless, if one or more of these components dominate, then it should be possible to significantly improve the scaling of serial TAD over intermediate length-scales by improving the scaling of these dominant components. We now consider the actual scaling behavior obtained in simulations of Ag/Ag(100) growth and Ag/Ag(100) annealing.

III. SCALING RESULTS FOR SERIAL TAD

In order to determine the scaling behavior of serial TAD, we have carried out serial TAD simulations of Ag/Ag(100) thin-film growth and annealing. In our simulations, an EAM (embedded atom method) potential for Ag developed by Adams *et al.*[49] was used. We note that this potential yields activation barriers for Ag diffusion processes on the Ag(100) surface which are close to those obtained from DFT calculations.[50] In both our growth and annealing simulations, the substrate was composed of five layers with the top two moving (2M) and bottom three fixed (3F) while periodic boundary conditions (PBCs) were applied parallel to the substrate. In order to study the scaling as a function of N, the lateral substrate size L was varied from $L = 6a$ to a maximum value of $L = 12a$ ($16a$) in the case of growth (annealing), where $a = 4.09$ Å is the lattice constant for Ag. For simplicity, our system size was measured in units of the smallest initial substrate size, e.g., $6a \times 6a$ corresponds to unit system-size $N_s = 1$ while a larger $12a \times 12a$ substrate corresponds to $N_s = 4$.

In our growth simulations, one monolayer (ML) of Ag was deposited on an initial Ag(100) substrate at a temperature $T_{\mathrm{low}} = 100$ K ($T_{\mathrm{high}} = 550$ K) with deposition rate $F = 5$ ML/s while an initial kinetic energy of 0.2 eV was used for the depositing atoms. In contrast, our annealing simulations were carried out at 3 different fixed coverages using starting configurations obtained from our growth simulations. One set of annealing simulations was carried out at $\theta = 0.05$ ML for $t = 10$ s with $T_{\mathrm{low}} = 150$ K ($T_{\mathrm{high}} = 750$ K) while additional annealing simulations were carried out at $\theta = 0.2$ and 1 ML for $t = 0.1$ s at $T_{\mathrm{low}} = 100$ K ($T_{\mathrm{high}} = 600$ K). In all of our simulations the high temperature used corresponded to the optimum value obtained in previous TAD simulations,[13] while an uncertainty $\delta = 0.1$ was used. Since our primary goal is to understand the scaling behavior, a relatively high value of the minimum prefactor ($\nu_{\mathrm{min}} = 10^{12}$ s^{-1}) was used in order to speed up our simulations. Our high-temperature MD simulations were carried out using a Langevin thermostat[51] with a friction coefficient of 10^{12} s^{-1} and a time step of 4 fs.

In most of our TAD simulations, one MD block consisted of $n_{\mathrm{MD}} = 500$ MD steps (corresponding to 2 ps) followed by transition detection. For the purposes of transition refining, e.g., to identify the first transition during each MD block, high-temperature MD configurations were saved every MD step. However, unless finer refining was needed (see below) only the configurations which were saved every $n_{\mathrm{MD}}^{\mathrm{sub}} = 5$ MD steps (corresponding to a sub-block) were used. In our tests of the performance of our localized refining (LREF) and localized saddle-point (LSAD) methods, we have also used a smaller block size of $n_{\mathrm{MD}} = 250$ along with $n_{\mathrm{MD}}^{\mathrm{sub}} = 10$. Energy barriers for new attempted events were calculated using the cNEB method with 11 images. All of our TAD growth and annealing simulations were carried out using the 2.6 GHz

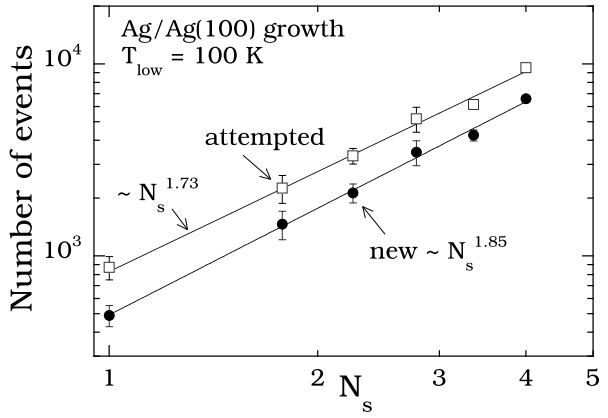

FIG. 2. Number of new and attempted events as function of N_s observed in TAD simulations of 1 ML of Ag/Ag(100) growth at $T_{low} = 100$ K with deposition rate $F = 5$ ML/s and $T_{high} = 550$ K.

IBM 1350 Glenn cluster at the Ohio Supercomputing Center (OSC).

A. Ag/Ag(100) growth

Before presenting our results for the overall scaling of serial TAD during the growth of 1 ML of Ag/Ag(100), we first consider the dependence of the number of high-temperature attempted events n_{att} on system size N_s. We note that here n_{att} only includes the first attempted event in each block and does not include all subsequent attempted events within the same block. As can be seen in Fig. 2, the total number of attempted events n_{att} and the number of new attempted events n_{new} scale as $N_s^{1.73}$ and $N_s^{1.85}$, respectively, in good agreement with the prediction $N_s^{2-\gamma} = N_s^{1.82}$. In addition, as expected, the total number of accepted events (not shown) scales linearly with N_s. Interestingly, approximately one-half of the attempted events are "repeat" events, e.g., the same attempted event occurs more than once before an event is accepted, although the fraction decreases slightly with increasing system-size. As shown in Fig. 3, the activation barriers for attempted events

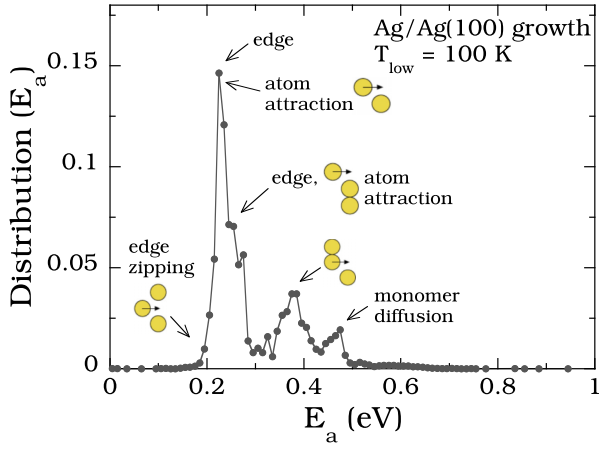

FIG. 3. Distribution of activation energies obtained in TAD simulations of Ag/Ag(100) growth. Labels indicate different types of events including edge-zipping, atom attraction, and edge and monomer diffusions.

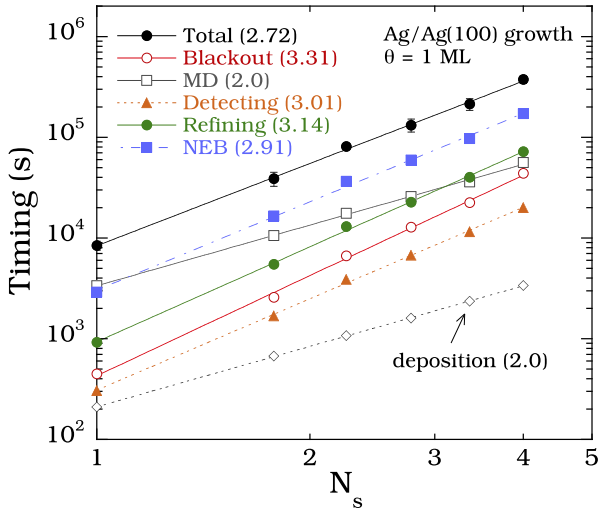

FIG. 4. Scaling of total TAD execution time along with times for individual components as a function of system-size N_s. Timings were obtained in serial TAD simulations of 1 ML Ag/Ag(100) growth without localization at $T_{low} = 100$ K and $T_{high} = 550$ K with flux $F = 5$ ML/s. Scaling exponents obtained from power-law fits are shown in parentheses.

span a wide range of energies from 0.05 eV to 1.08 eV and include a variety of processes including edge-zipping, single-bond edge-diffusion, atom attraction, and monomer hopping. However, most of the events correspond to barriers ranging from 0.15 eV to 0.7 eV.

We now consider the scaling of the individual components of serial TAD as well as the overall scaling. As can be seen in Fig. 4, the dominant contribution to the TAD timing is the cost of NEB calculations (filled squares) which scales approximately as $N_s^{2.91\pm0.18}$, in reasonable agreement with the prediction $N_s^{3-\gamma} = N_s^{2.82}$. For the largest system-size ($N_s = 4$) the next dominant contribution corresponds to transition refining (filled circles) which scales as $N_s^{3.14}$. These contributions are then followed (in order of decreasing importance for $N_s = 4$) by high-temperature MD ($N_s^{2.0}$), blackout ($N_s^{3.31}$), transition detection ($N_s^{3.01}$), and deposition ($N_s^{2.0}$). We note that the exponents describing the scaling of all 5 components of TAD, e.g., NEB, transition refining, MD, blackout, and transition detection are somewhat higher (and in the case of blackout significantly higher) than the expected $N_s^{3-\gamma}$ ($N_s^{2-\gamma}$ for MD) behavior. However, because the exponents associated with both MD and deposition are both significantly lower than $3 - \gamma$, the overall exponent ($\alpha \simeq 2.72$) associated with the *total* serial TAD computational time is also slightly lower.

We conjecture that the somewhat higher-than-expected value of the exponent for MD (e.g., 2 rather than $2 - \gamma$) is related to the fact that—in contrast to an ideal TAD calculation—we do not immediately detect a transition and stop the high-temperature MD as soon as an event occurs. As a result, the MD time during a given block which occurs after the first attempted event is wasted. Since the probability of the first attempted event occurring at the beginning of a block increases with system size, this implies a somewhat increased scaling exponent compared to the theoretical prediction. This is also consistent with the observation that the total number of

attempted events during a block—including *all* attempted events which occur after the first attempted event—is approximately proportional to N_s^2. It is also confirmed by results we have obtained in the case of annealing (see below) in which we have measured the amount of wasted MD time as a function of system-size. We note that, assuming a fixed block size, this scaling behavior is also expected in the limit of large N_s since in this limit the high-T MD time required to accept an event becomes small compared to the block size, while the actual MD time is equal to the block-size. Since $t_{low,short}$ scales as $1/N_s$, for fixed block size the ratio of the high-temperature MD time to low-temperature time scales as N_s which implies that the computational work associated with MD scales as N_s^2, in good agreement with our results.

The somewhat higher than expected scaling exponent for transition detection (3.0) is also consistent with that obtained for high-temperature MD since the number of blocks is proportional to the high-temperature MD time while the cost per minimization is proportional to N_s. In contrast, the exponent for transition refining is somewhat higher (3.14) than that associated with transition detection. This may be related to the fact that since the number of attempted events for a given block size increases with system-size, the probability of a transition occurring early in a given block also increases linearly with system-size, thus requiring logarithmically more bisections (whose cost scales as N_s) with increasing system size. In this connection, we find that for a fixed system-size the number of bisections $n_{\rm bis}$ performed per refining process is given by $n_{\rm bis} \sim \ln(n_{\rm MD}/n_{\rm MD}^{\rm sub})$.

Fig. 4 also indicates that the process of blackout or decorrelation, which occurs after any attempted transitions during the MD, deviates from the expected $N_s^{3-\gamma} \sim N_s^{2.82}$ scaling behavior and instead scales as $N_s^{3.31}$. The relatively large increase in the scaling exponent in this case is due to the possibility of additional transitions during the blackout period, which may necessitate multiple executions of the blackout process (see Fig. 1). Since the probability of additional transitions during the blackout period increases with N_s, this increases the scaling exponent for this process.

B. Ag/Ag(100) annealing

We have also examined the scaling of serial TAD in the case of Ag/Ag(100) annealing at $T_{\rm low} = 150$ K (with $T_{\rm high} = 750$ K and coverage $\theta = 0.05$ ML) and at $T_{\rm low} = 100$ K (with $T_{\rm high} = 600$ K and $\theta = 0.2$ ML and 1 ML). Here we note that the high temperatures $T_{\rm high}$ were chosen in such a way as to be the optimal high temperatures for the dominant diffusion processes in the annealing systems, as suggested in Ref. 13. For $\theta = 0.2$ and 1.0 ML, edge diffusion and atom attraction with $E_a = 0.23$–0.26 eV are the dominant processes, while for $\theta = 0.05$ ML, monomer diffusion with $E_a = 0.48$ eV is dominant.

Figure 5 shows the detailed scaling behavior in this case. As in Fig. 2, in Fig. 5(a) we include only the first attempted event in each block since subsequent events are ignored at the end of the transition refining process. As can be seen, the number of "new" or previously unseen attempted

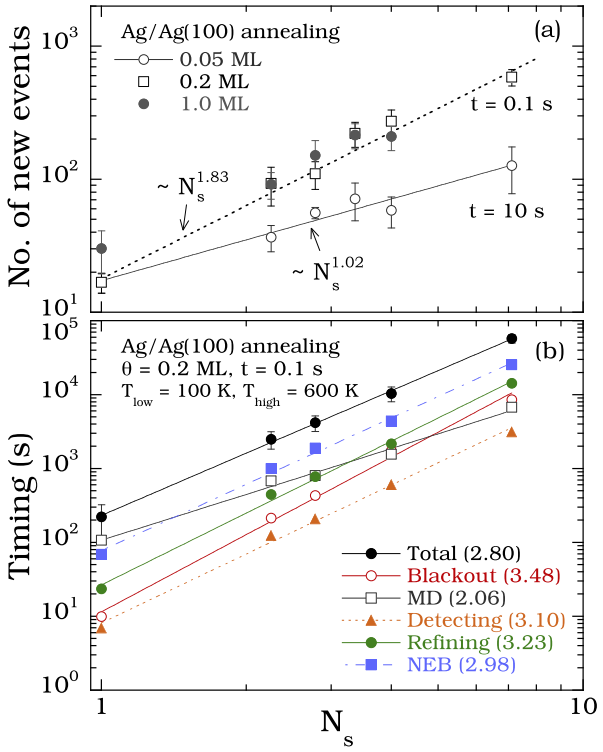

FIG. 5. (a) Number of new events observed in TAD simulations of Ag/Ag(100) annealing for $t = 10$ s ($\theta = 0.05$ ML) and for $t = 0.1$ s ($\theta = 0.2$ and 1 ML). $T_{\rm low} = 150$ K and $T_{\rm high} = 750$ K for $\theta = 0.05$ ML annealing while $T_{\rm low} = 100$ K and $T_{\rm high} = 600$ K for both $\theta = 0.2$ and 1 ML annealing simulations. (b) Total TAD execution time along with times for individual components as function of N_s obtained in serial TAD simulations of Ag/Ag(100) annealing at $\theta = 0.2$ ML for $t = 0.1$ s without localization.

events observed at higher coverages ($\theta = 0.2$ and 1.0 ML) exhibits the expected scaling behavior $n_{\rm new} \sim N_s^{2-\gamma} \sim N_s^{1.83}$ (where $\gamma = 1/6$) with a broad distribution of energy barriers for attempted events (0.17 – 1.02 eV). On the other hand, at low coverage ($\theta = 0.05$ ML) the number of new events increases linearly with N_s. This is due to the simple nature of the configurations in this case for which approximately 80% of attempted events correspond to monomer diffusion while the number of monomers in the system is simply proportional to N_s.

The scaling of the total computational time as a function of system-size during TAD annealing at $\theta = 0.2$ ML is shown in Fig. 5(b). As can be seen, the overall scaling exponent $\alpha \simeq 2.80 \pm 0.32$, is very similar to that obtained for Ag/Ag(100) growth and is in reasonable agreement with the expectation that $\alpha \simeq 3 - \gamma \simeq 2.83$. As for the case of growth, the (NEB) saddle-point calculation and the process of refining are the two dominant components for larger system-sizes. Similarly, the computational time for MD (transition detection) scales approximately as N_s^2 (N_s^3) rather than as $N_s^{2-\gamma}$ ($N_s^{3-\gamma}$). However, perhaps due in part to the somewhat higher value of $T_{\rm high}$, the scaling exponents for blackout (3.48), NEB (2.98), and transition refining (3.23) are slightly higher than in the case of growth.

We note that, just as for the case of Ag/Ag(100) growth, the discrepancies between the theoretical predictions for the scaling of MD, transition detection, and transition refining are

due in part to the existence of wasted MD time between the first event in a block and the end of the block. To verify that this is the case, we have separately measured the contributions t_1 (t_w) to the total MD time corresponding to the sum of MD times in each block before (after) the first attempted event and have found that while t_1 scales as $N_s^{1-\gamma}$ as expected, the amount of wasted time t_w scales as $N_s^{1.4}$. As a result, the total MD time scales effectively as N_s while the total computational work scales as N_s^2.

IV. USING LOCALIZATION TO IMPROVE THE SCALING OF SERIAL TAD

Our scaling results for Ag/Ag(100) growth and Ag/Ag(100) annealing indicate that for moderate-to-large system sizes, the two dominant contributions to the computational cost of TAD are (a) the calculation of activation energies for attempted events via NEB and (b) the process of transition refining. Accordingly, to improve the scaling of serial TAD with system size, we have developed and implemented methods to localize both of these calculations. In addition, we have considered the use of on-the-fly transition detection to slightly further improve the scaling of serial TAD.

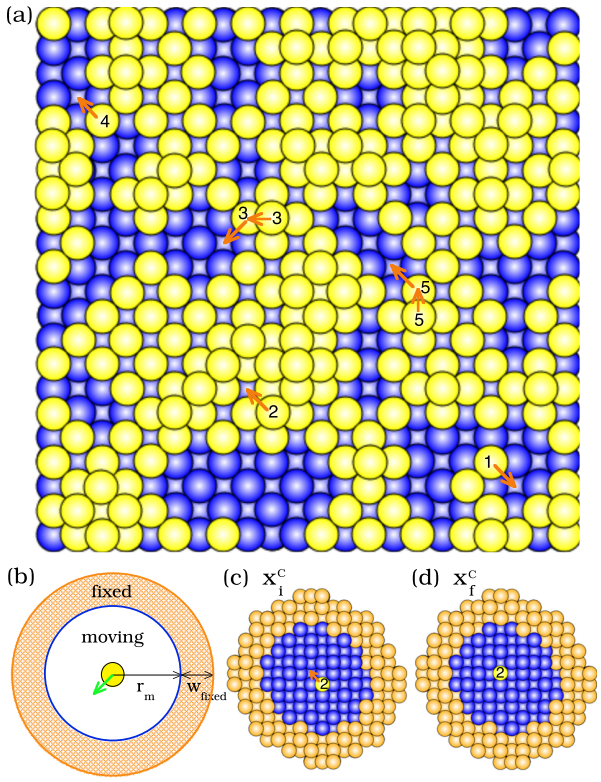

FIG. 6. (a) Five attempted transitions observed after a 2 ps MD run in a TAD annealing simulation of Ag/Ag(100) at $T_{\text{high}} = 600$ K for system size $N_s = (11/6)^2$. Here, the activation barriers for transitions 1–5 are 0.240 eV, 0.239 eV, 0.50 eV, 0.249 eV, and 0.385 eV, respectively. Event 2 with $E_a = 0.239$ eV is found to be the first transition after refining process. (b) Schematic diagram of LREF method describing moving and fixed regions around a monomer (yellow atom). (c) and (d) are top views of the chucks constructed for the atom 2 in (a) for the initial and final states, respectively, where blue and orange atoms represent moving and fixed atoms, respectively.

A. Local saddle-point calculations

Recently, we have developed a localized saddle-point (LSAD) method[52] which instead of using the whole system, uses only a small region around the transition to calculate the corresponding activation barrier. In addition, we have shown that for a fixed event size, this method leads to system-size independent scaling, while maintaining a negligibly small error. As a result, we expect that the use of our LSAD method in TAD simulations will improve the scaling for this portion of TAD by a factor of N.

B. Local transition refining

We now discuss the use of localization to improve the scaling of transition refining, which is carried out after a new transition is detected via global minimization at the end of a block. In serial TAD this is normally done using a series of global minimizations which are carried out using bisection. In particular, once a transition is detected at the end of a block, the saved configuration at the middle of the block is minimized. If this minimized configuration differs from (is the same as) the initial state, then an additional global minimization is performed for a configuration corresponding to an earlier (later) time. This process continues until two successive saved states have been identified such that one corresponds to the initial state and the other corresponds to a transition.

In order to improve the scaling of this refining process with system-size, we have developed a local refining method (LREF) in which global minimizations are replaced whenever possible by local minimizations. For example, if a single transition is identified to have occurred at the end of a block, then all subsequent minimizations during the process of refining are only performed on a chunk of atoms surrounding the transition region. In particular, in the case of our Ag/Ag(100) simulations, once all the atoms involved in the transition have been identified, then in all subsequent minimizations all atoms which are within a distance $r_m = 9.0$ Å (approximately the 5th neighbor distance) from either the initial or the final position of all of the atoms involved in the transition are included as moving atoms in the local "chunk" which is used for minimization, while an additional distance w_{fixed} corresponding to the EAM potential cutoff distance is used to surround the moving atoms with fixed atoms (see Fig. 6(b)).

If there is no more than one transition in the system at the end of a block we expect that this method will improve the scaling of transition refining by a factor of N. However, with increasing system size there is an increasing probability that two or more distinct transitions may have occurred at the end of a block (see Fig. 6(a)). In order to detect this possibility, we apply a nearest-neighbor cluster-counting algorithm. In particular, all atoms involved in transitions (e.g., all atoms whose minimized positions have been displaced from the initial minimized configuration by an amount greater than $d_{\text{min}} \simeq 0.8$ Å) which are also within a distance of $1.05\ a_1$ from one another (where a_1 is the nearest-neighbor distance) are assumed to belong to the same transition. Each connected

cluster of displaced atoms is then assumed to correspond to a distinct transition, each of which may have occurred at different times during the high-temperature MD block. In order to determine the first transition in the block, a series of bisections and local minimizations is carried out for each distinct transition.

In most cases the local bisections and minimizations were carried out using the high-temperature configurations which were saved every sub-block, e.g., every $n_{\mathrm{MD}}^{\mathrm{sub}}$ MD steps. However, in some cases, especially for larger system-sizes, more than one transition may occur between successive saved sub-block configurations. If this occurs, then a finer refining process is carried out based on the configurations which are saved after every single MD step. In the relatively rare case that multiple distinct transitions correspond to the first identified transitions in an MD block, one of these transitions is selected randomly. However, in some cases multiple distinct transitions may occur which are both close in space and in time, and in these cases our localized method may fail. In this case, global minimization is used.

C. On-the-fly transition detection

In addition to the use of localized saddle-point calculations (LSAD) and localized transition refining (LREF) we have also explored the use of on-the-fly transition detection to further improve the scaling of serial TAD. In particular, we have considered the use of bond-stretching and/or atom displacement combined with local minimization in order to quickly identify possible transitions during the high-temperature MD. We note that such a local on-the-fly transition search is expected to be effective if there is a well-defined critical parameter characterizing a transition, such as the atom displacement or bond-stretching at the saddle point.

As an example, in Ag/Ag(100) annealing at very low coverage, monomer diffusion is the main activated process while atom attraction and edge-diffusion also play a role. In particular, the in-plane displacement for monomer hopping at the saddle point $d_c^{\mathrm{sad}} = a_1/2$, where $a_1 \simeq 2.89$ Å is the nearest-neighbor bond-length while the saddle-point displacements for edge-diffusion and atom attraction are approximately the same. In this case we have found that on-the-fly detection of an activated event can be carried out very efficiently by locally minimizing the configuration whenever the high-temperature lateral monomer displacement d is greater than half the nearest-neighbor distance.

However, for the slightly more complex example of Ag/Ag(100) annealing at coverage $\theta = 0.2$ ML, the value of the maximum saddle point displacement d_c^{sad} calculated for a variety of transitions has a relatively broad distribution with 0.8 Å $\leq d_c^{\mathrm{sad}} \leq$ 2.0 Å. In addition, we have found that it is strongly correlated with the value of the energy barrier E_a (roughly, $d_c^{\mathrm{sad}} \sim E_a^{1/3}$ for $E_a \leq 0.8$ eV) which is not yet determined at this stage of refining process. A similarly broad distribution was also obtained for Cu/Cu(100) as well as for the case in which the bond-length was used instead of displacement. As a result, it is hard to determine when to minimize those possible candidates from complex high-temperature MD configurations without significantly sacrificing efficiency.

Accordingly, based on our knowledge of the Ag/Ag(100) system we have used a more limited on-the fly approach to slightly enhance the speed of our simulations. In particular, when the displacement d during a high-temperature MD block was found to never exceed $d_{\mathrm{min}} = 0.8$ Å, transition detection and refining were not carried out. Conversely, if at some point during a block it was found that $d > d_{\mathrm{max}}$, where $d_{max} = 2.0$–2.2 Å, then after the block was completed the refining process was only carried out over an interval between the beginning of the block and the corresponding sub-block, thus reducing the number of bisection iterations during the refining process. Since both of these conditions are relatively rare, this led to a relatively small (e.g., 1%–2%) improvement of the simulation speed.

D. Results

Fig. 7 shows the scaling behavior obtained in our serial TAD simulations using localization (LTAD) for the case of Ag/Ag(100) growth. As can be seen, the scaling of the saddle-point (NEB) calculations has been reduced by a factor of N_s compared to ordinary serial TAD, e.g., the corresponding exponent was reduced from a value of 2.9 in the absence of localization to 1.8 with localization. Similarly, the exponent corresponding to transition refining was reduced from a value of 3.2 in the absence of localization to 2.2 with localization, leading to an overall factor of N_s improvement in the scaling behavior for transition refining. However, since the scaling of the other components of TAD including MD, blackout, and transition detection was essentially unchanged, the overall scaling of TAD was improved by a somewhat smaller factor of $N^{1/2}$. In particular, the exponent characterizing the overall scaling has been reduced from $\alpha \simeq 2.7$ in the case of the usual serial TAD to $\alpha \simeq 2.2$ by using localization. We note that for the largest system size ($N_s = 4$) this leads to a reduction of

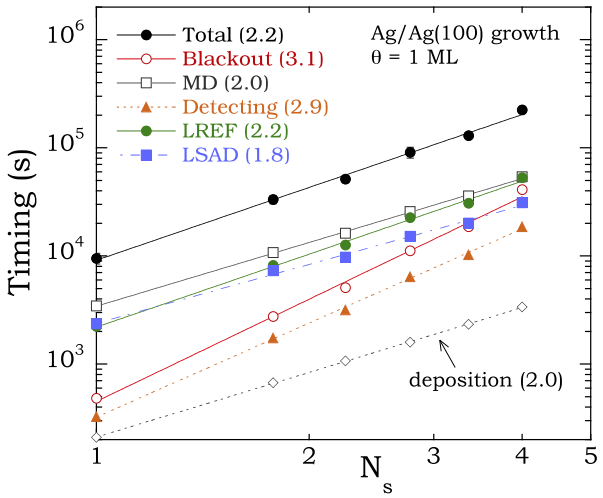

FIG. 7. Scaling of total TAD execution time along with times for individual components as a function of system size obtained in serial TAD simulations of Ag/Ag(100) growth with localization (LREF and LSAD methods). Scaling exponents obtained from power-law fits are shown in parentheses.

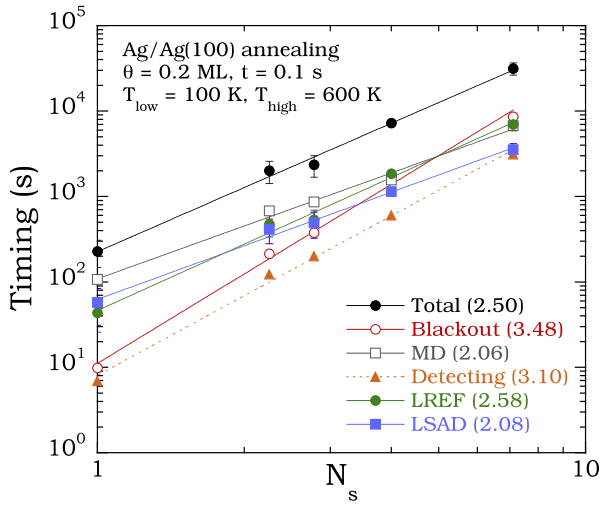

FIG. 8. Scaling of total TAD execution time along with times for individual components as a function of system size obtained in serial TAD simulations of Ag/Ag(100) annealing with localization (LREF and LSAD methods). Scaling exponents obtained from power-law fits are shown in parentheses.

approximately 40% in the total TAD simulation time compared to ordinary serial TAD.

Similar results are shown in Fig. 8 for the case of Ag/Ag(100) annealing at coverage $\theta = 0.2$ ML for 0.1 s. In this case, the scaling of the saddle-point calculations has been reduced via localization by a factor of approximately $N_s^{0.9}$, with the corresponding exponent reduced from a value of 2.98 in the absence of localization to 2.08 with localization. Similarly, the scaling of the transition refining calculations has been reduced by a factor of approximately $N_s^{0.65}$, with the corresponding exponent reduced from a value of 3.23 in the absence of localization to 2.58 with localization. As a result, the overall scaling of TAD annealing was improved by a factor of approximately $N^{0.3}$, e.g., the exponent characterizing the overall scaling was reduced from $\alpha \simeq 2.8$ in the case of the usual serial TAD to $\alpha \simeq 2.5$ by using localization. For the largest system size used in this case ($N_s = 7$) this leads to a reduction of approximately 55% in the total TAD simulation time. As shown in Table I, this corresponds to an overall performance improvement compared to serial TAD annealing simulations without localization of approximately 1.8–1.9.

TABLE I. Performance P of LTAD (corresponding to ratio of time for ordinary serial TAD to that for LTAD) in Ag/Ag(100) annealing at $\theta = 0.2$ ML for $t = 0.1$ s with $N_s \simeq 7.11$, $T_{low} = 100$ K, and $T_{high} = 600$ K. P_{LREF} corresponds to performance using only localized refining, while $P_{LREF+LSAD}$ corresponds to both localized refining and localized saddle-point calculations. The subscript fref represents finer refining (see text).

n_{MD}/n_{MD}^{sub}		t_{global} (h)	P_{LREF}	$P_{LREF+LSAD}$
500/20	25	15.2 ± 2.5	1.05	1.81
500/10	50	15.9 ± 3.0	1.17	1.86
250/10	25	16.3 ± 1.7	1.14	1.80
250/5	50	17.5 ± 1.7	1.18	1.89
$500/20_{fref}$	25		1.11	1.82
$500/10_{fref}$	50		1.20	1.89
$250/10_{fref}$	25		1.20	1.88

We have also examined the effect of the MD block size n_{MD} and the ratio n_{MD}/n_{MD}^{sub} of the block size to the sub-block size on performance in the case of annealing. As shown in Table I, for the case of serial TAD without localization, the overall computational cost increases with decreasing block size n_{MD} since this leads to an increased number of minimizations at the end of the block for which no transition has occurred. Similarly, in this case decreasing the sub-block size while keeping the block-size fixed leads to an increased computational cost due to an increase in the number of refining steps. However, in the case of serial TAD with localization, these effects appear to be significantly smaller due to the decreased cost of refining.

Table I also includes performance results for Ag/Ag(100) annealing obtained by carrying out additional finer refining ("fref") when an intermediate state is expected after the first stage of refining. We note that a NEB calculation for the initial and final states which contains an intermediate state may spend a significant amount of computational time before either resolving or failing to resolve the intermediate state, resulting in a significant waste of computational time. In this finer refining process, we continue refining below the sub-block level, e.g., on the scale of every MD step, in order to find the first MD step at which the transition occurs. While this finer refining method cannot completely remove the possibility of intermediate states, since two events could occur at the same MD step, it noticeably increases the performance by speeding up the NEB calculations, as can be seen in Table I.

V. SUMMARY

In order to investigate the scaling behavior of serial TAD, we have carried out extensive calculations as a function of system-size for both the case of Ag/Ag(100) growth and Ag/Ag(100) annealing. Our results for the overall scaling behavior in the absence of localization are in reasonable agreement with theoretical predictions. However, we have also found that, due to a variety of factors discussed in Sec. IV, the scaling exponents for all of the key components of TAD, including NEB, blackout, MD, transition detection, and transition refining, are actually somewhat larger than expected from simple theoretical predictions.

Our results also indicate that, for moderate-to-large system sizes, the dominant contributions to the cost of TAD are saddle-point (NEB) calculations as well as transition refining. Accordingly, in order to improve the overall scaling as a function of system size, we have developed a localized saddle-point (LSAD) method to calculate the activation energies as well as a localized transition refining (LREF) method. Using these methods, we have found that the overall scaling of localized TAD (LTAD) in the case of Ag/Ag(100) growth for moderate-to-large system sizes is improved by approximately a factor of $N^{1/2}$ compared to "global" TAD, thus leading to an overall scaling exponent $\alpha \simeq 2.2$. Similarly, in the case of Ag/Ag(100) annealing, the overall scaling of LTAD was improved by a somewhat smaller overall factor of approximately $N^{0.3}$, thus leading to an overall exponent $\alpha \simeq 2.5$. One of the possible reasons for the somewhat smaller

improvement in the overall scaling behavior in the case of Ag annealing compared to growth is the use of a higher T_{high} in this case which leads to a larger number of attempted events per block.

We have also investigated the possibility of using on-the-fly transition detection based on individual atomic displacements and/or bond stretching or compression in order to further improve the scaling of LTAD. However, we find that even for simple homoepitaxial systems such as the Ag/Ag(100) system studied here, the distribution of saddle-point displacements is very broad. As a result, it is hard to develop an efficient criterion for detecting transitions on-the-fly. However, by using very conservative, system-specific criteria we were able to obtain a small additional further improvement, e.g., approximately 1%–2%, in the simulation speed. In the future, it would be of interest to try to develop a more general on-the-fly transition detection method in order to further improve the scaling of TAD.

Our simulations also indicate that one component of TAD which is expected to become particularly significant for very large system-sizes is blackout, since it exhibits particularly poor scaling behavior. One of the primary reasons for this poor scaling behavior, which corresponds to an exponent which is significantly larger than $3 - \gamma$, is that the total rate of basin escape during the blackout period increases linearly with increasing system size. As a result, the probability that multiple iterations of high-temperature MD followed by minimization are required during each blackout period increases with system-size. Therefore, it would be of interest in future work to develop methods to improve and or mitigate the scaling of this component of TAD.

Finally, we note that while here we have focused exclusively on the development of serial methods based on localization to improve the scaling of serial TAD, by using parallel methods such as our previously developed parTAD algorithm,[42] it should be possible to further improve the scaling behavior. For example, by carrying out parallel MD on the entire system, the scaling of the other 3 components of TAD, e.g., high-temperature MD, blackout, and transition detection, can be improved by a factor of N. In addition, the use of parallel NEB should lead to an additional speed-up of our localized NEB (LSAD) calculations by a factor corresponding to the number of images. Similarly, the use of multiple TAD replicas for a given basin can be used[3,19] to speed-up the search for accepted transitions. As a result, we expect that by combining our localization techniques with parallelization, it should be possible to further improve the speed of TAD calculations for moderate-to-large system-sizes.

ACKNOWLEDGMENTS

This work was supported by NSF Grant No. DMR-1410840 as well as by a grant of computer time from the Ohio Supercomputer Center.

[1]A. F. Voter, F. Montalenti, and T. C. Germann, Annu. Rev. Mater. Res. **32**, 321 (2002).
[2]D. Perez, B. P. Uberuaga, Y. Shim, J. G. Amar, and A. F. Voter, Annu. Rep. Comput. Chem. **5**, 79 (2009).
[3]D. Perez, B. P. Uberuaga, and A. F. Voter, Comput. Mater. Sci. **100**, 90 (2015).
[4]A. F. Voter, J. Chem. Phys. **106**, 4665 (1997).
[5]A. F. Voter, Phys. Rev. Lett. **78**, 3908 (1997).
[6]A. F. Voter, Phys. Rev. B **57**, R13985 (1998).
[7]B. P. Uberuaga, S. J. Stuart, and A. F. Voter, Phys. Rev. B **75**, 014301 (2007).
[8]C.-Y. Lu, A. F. Voter, and D. Perez, J. Chem. Phys. **140**, 044116 (2014).
[9]R. J. Zamora, B. P. Uberuaga, D. Perez, and A. F. Voter, Annu. Rev. Chem. Biomol. Eng. **7**, 87 (2016).
[10]M. R. Sørensen and A. F. Voter, J. Chem. Phys. **112**, 9599 (2000).
[11]F. Montalenti, M. R. Sørensen, and A. F. Voter, Phys. Rev. Lett. **87**, 126101 (2001).
[12]F. Montalenti and A. F. Voter, J. Chem. Phys. **116**, 4819 (2002).
[13]Y. Shim and J. G. Amar, J. Chem. Phys. **134**, 054127 (2011).
[14]R. A. Miron and K. A. Fichthorn, Phys. Rev. Lett. **93**, 128301 (2004).
[15]K. A. Fichthorn, R. A. Miron, Y. Wang, and Y. Tiwary, J. Phys. Condens. Matter **21**, 084212 (2009).
[16]K. A. Fichthorn and S. Mubun, Comput. Mater. Sci. **100**, 104 (2015).
[17]S. Hara and J. Li, Phys. Rev. B **82**, 184114 (2010).
[18]A. Ishii, S. Ogata, H. Kimizuka, and J. Li, Phys. Rev. B **85**, 064303 (2012).
[19]S. Divi and A. Chatterjee, J. Chem. Phys. **140**, 184115 (2014).
[20]V. Bochenkov, N. Suetin, and S. Shankar, J. Chem. Phys. **141**, 094105 (2014).
[21]A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).
[22]A. Laio and F. L. Gervasio, Rep. Prog. Phys. **71**, 126601 (2008).
[23]A. Kushima, X. Lin, J. Li, J. Eapen, J. C. Mauro, X. Qian, P. Diep, and S. Yip, J. Chem. Phys. **130**, 224504 (2009).
[24]B. M. Dickson, Phys. Rev. E **84**, 037701 (2011).
[25]C.-Y. Lu, D. E. Makarov, and G. Henkelman, J. Chem. Phys. **133**, 201101 (2010).
[26]G. T. Barkema and N. Mousseau, Phys. Rev. Lett. **72**, 4358 (1996).
[27]L. Munro and D. J. Wales, Phys. Rev. B **59**, 3969 (1999).
[28]G. Henkelman and H. Jónnson, J. Chem. Phys. **111**, 7010 (1999); **113**, 9978 (2000).
[29]N. Mousseau, L. K. Béland, P. Brommer, F. El-Mellouhi, J.-F. Joly, G. K. N'Tsouaglo, O. Restrepo, and M. Trochet, Comput. Mater. Sci. **100**, 111 (2015).
[30]A. Heyden, A. T. Bell, and F. J. Keil, J. Chem. Phys. **123**, 224101 (2005).
[31]Y. Fan, S. Yip, and B. Yildiz, J. Phys.: Condens. Matter **26**, 365402 (2014).
[32]P. Xiao, J. Duncan, L. Zhang, and G. Henkelman, J. Chem. Phys. **143**, 244104 (2015).
[33]J. A. Sprague, F. Montalenti, B. P. Uberuaga, J. D. Kress, and A. F. Voter, Phys. Rev. B **66**, 205415 (2002).
[34]B. P. Uberuaga, R. Smith, A. R. Cleave, F. Montalenti, G. Henkelman, R. W. Grimes, A. F. Voter, and K. E. Sickafus, Phys. Rev. Lett. **92**, 115505 (2004).
[35]M. Cogoni, A. Mattoni, B. P. Uberuaga, A. F. Voter, and L. Colombo, Appl. Phys. Lett. **87**, 191912 (2005).
[36]Y. Shim, V. Borovikov, B. P. Uberuaga, A. F. Voter, and J. G. Amar, Phys. Rev. Lett. **101**, 116101 (2008).
[37]X.-M. Bai, A. F. Voter, R. G. Hoagland, M. Nastasi, and B. P. Uberuaga, Science **327**, 1631 (2010).
[38]Y. Shim and J. G. Amar, Phys. Rev. B **81**, 045416 (2010).
[39]Y. Shim and J. G. Amar, Phys. Rev. Lett. **108**, 076102 (2012).
[40]G. H. Vineyard, J. Phys. Chem. Solids **3**, 121 (1957).
[41]S. Goedecker, F. Lançon, and T. Deutsch, Phys. Rev. B **64**, 161102(R) (2001).
[42]Y. Shim, J. G. Amar, B. P. Uberuaga, and A. F. Voter, Phys. Rev. B **76**, 205439 (2007).
[43]Y. Shim and J. G. Amar, Phys. Rev. B **71**, 125432 (2005).
[44]C. E. Botez, W. C. Elliott, P. F. Miceli, and P. W. Stephens, Phys. Rev. B **66**, 075418 (2002).
[45]C. E. Botez, P. F. Miceli, and P. W. Stephens, Phys. Rev. B **66**, 195413 (2002).
[46]C. E. Botez, K. Li, E. D. Lu, W. C. Elliott, P. F. Miceli, E. H. Conrad, and P. W. Stephens, Appl. Phys. Lett. **81**, 4718 (2002).
[47]E. Martínez, B. P. Uberuaga, and A. F. Voter, Phys. Rev. E **89**, 063308 (2014).
[48]G. Henkelman, B. P. Uberuaga, and H. Jónsson, J. Chem. Phys. **113**, 9901 (2000).
[49]J. B. Adams, S. M. Foiles, and W. G. Wolfer, J. Mater. Res. **4**, 102 (1989).
[50]Y. Shim and J. G. Amar, Phys. Rev. B **83**, 245419 (2011).
[51]M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, New York, 1987), p. 263.
[52]Y. Shim and J. G. Amar, J. Chem. Phys. **138**, 094101 (2013).