Going Through Rough Times:

Parallel Discrete Event Simulations (PDES)
--- A Physicist's Perspective

Mark A. Novotny

Dept. of Physics and Astronomy

HPC^2 Center for Computational Sciences

Funded in part the NSF

(DMR9871455) → DMR0113049 (ITR) → DMR0426488 (ITR)

ITMO 2012

Mississippi State UNIVERSITY

Collaborators

Alice Kolakowska (Purdue U, Calumet, former post-doc)

Gyorgy Korniss (Rensselaer Polytechnic Institute)

Zoltan Rácz (Eötvös University, Hungary)

Per Arne Rikvold (Florida State University)

Lev Shchur (Scientific Center in Chernogolovka, Russia)

Zoltan Toroczkai (Notre Dame University)

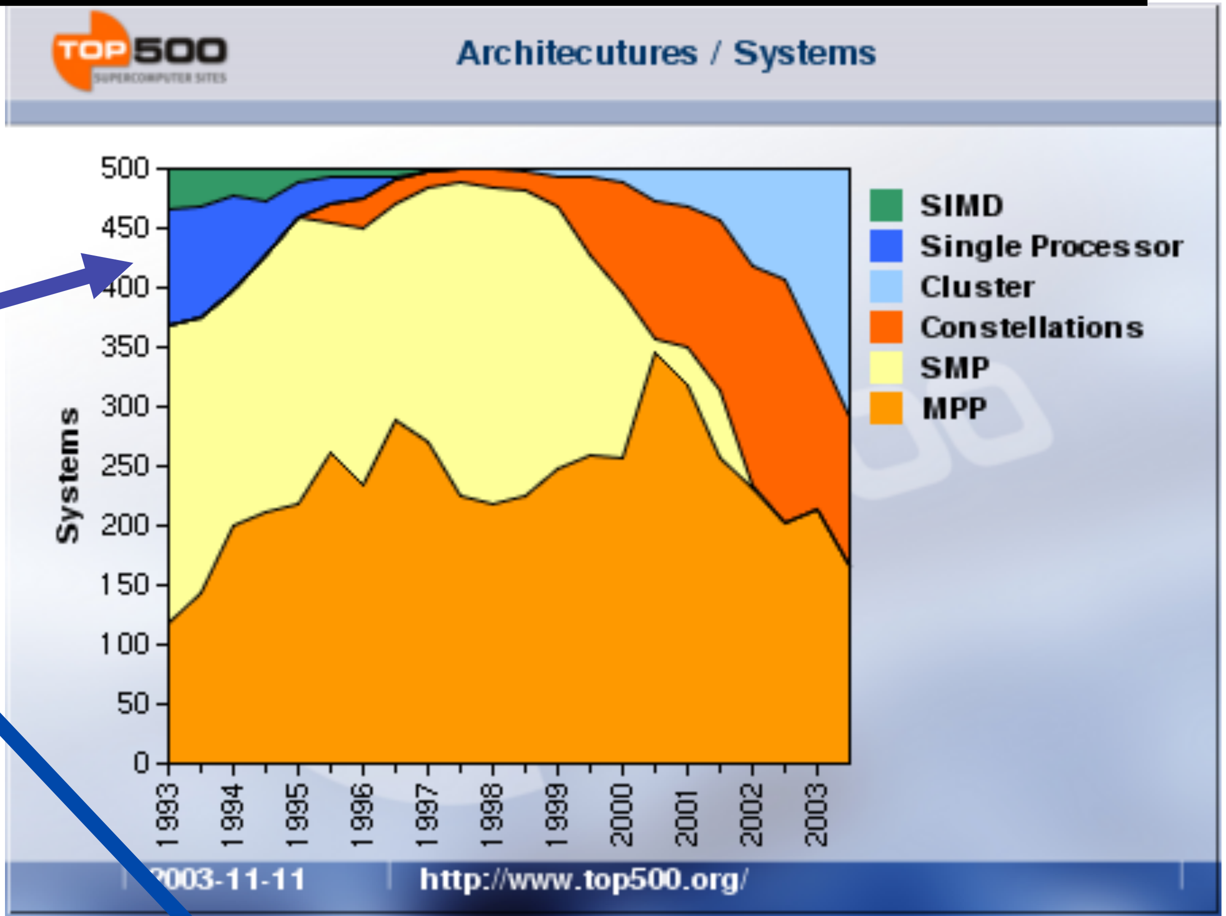
Students

H. Guclu, B. Kozma (Rensselaer Polytechnic Inst.)

Terrance Dubrues, Poonam Verma (Mississippi State University)

Daniel Brown, Melissa Cook, Sara Gill, Tori Norwood,
Amanda Novotny (summer High School students)

Problem: no signal faster than speed of light

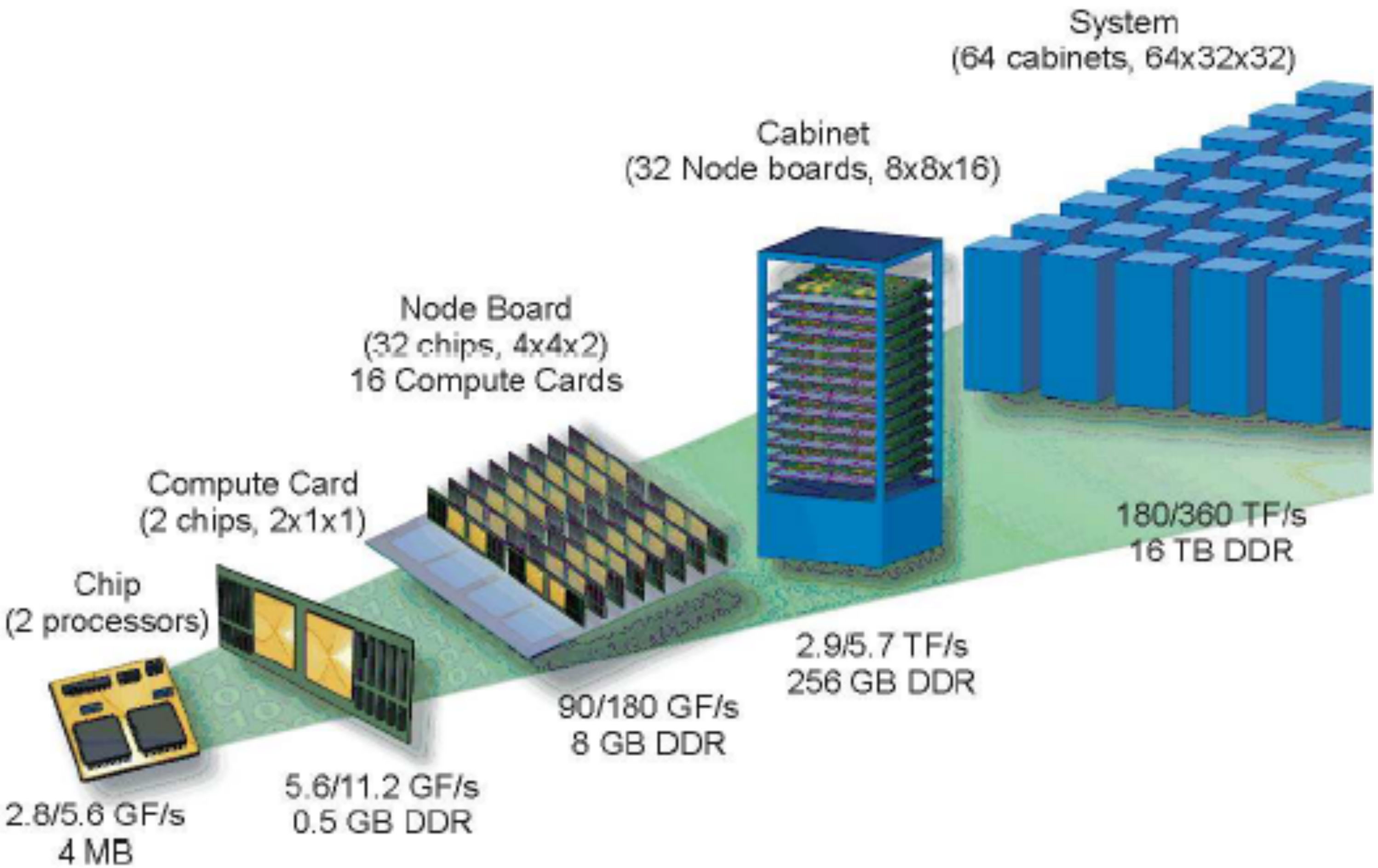


Since 1996 (at least)

computing on supercomputers *requires* parallel computing

Prediction: *in 10 years almost all PCs will have >16 PEs (Processing Elements)*

BlueGene/L → BlueGene/Q



2x2x32x32x64 = 131,072 PEs

June 2012 Top500.org

Sequoia computer
BlueGene/Q IBM
US DOE NNSA/LLNL
1,572,864 cores
R_{peak}=15.7 petaFLOP
R_{max}=20.1 petaFLOP

0 with less than 1024 Pes
1 Top-500 with 1024 to 2048 PEs

12 Top-500 with >=128,000 PEs

Perfect Scalability

N_{PE} = # Processing Elements

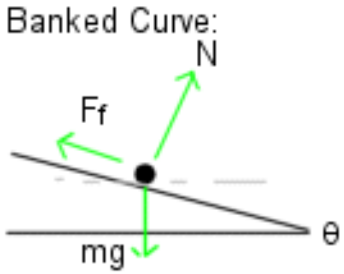
Non-trivial parallelization

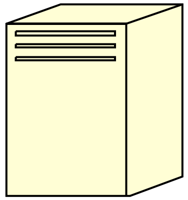
Amount of work $O(N_{PE})$

- For large N_{PE}, utilization independent of N_{PE}
- For large N_{PE}, PE memory independent of N_{PE}
- Number of interconnects $O(1)$ per PE

How Computational Physicists Count

1 body



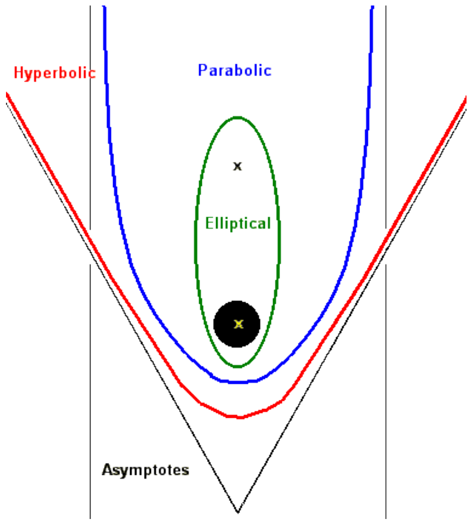


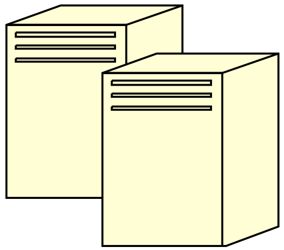
1 PE (Processor Element)

=

1 grad student

2 body





2 PEs

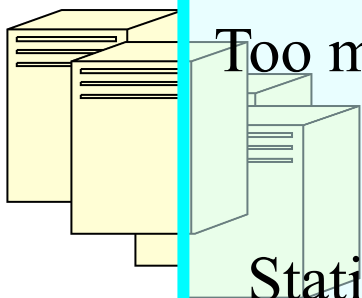
=

2 grad students

Too Many bodies

use

Statistical Mechanics



Too Many PEs

Too many grad students

use

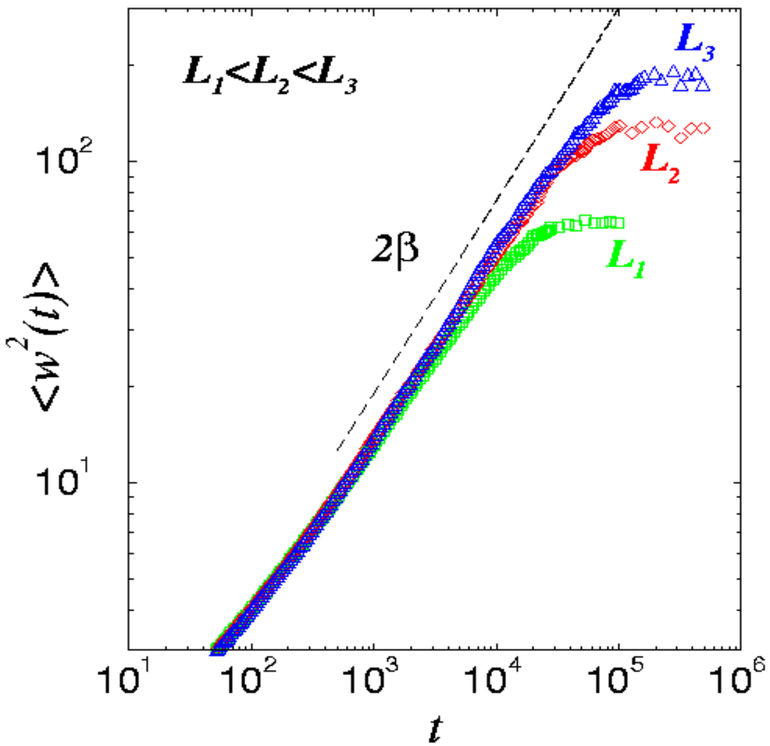
Statistical Mechanics

Complicated Behavior & Informatics
from Non-equilibrium Surface Growth Models

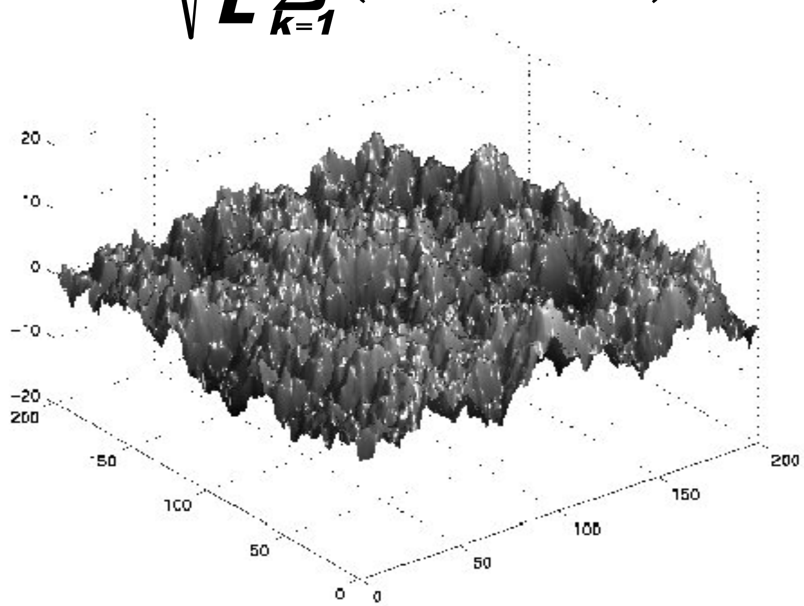
Motivation for **PDES** model

Parallel computing

Non-equilibrium surface growth



$$w(t)=\sqrt{\frac{1}{L}\sum_{k=1}^{L}\left(\tau_k(t)-\bar{\tau}(t)\right)^2}$$



$$\langle w^2(t)\rangle \sim \begin{cases} t^{2\beta}, & if \quad t \ll t_\times \\ L^{2\alpha}, & if \quad t \gg t_\times \end{cases}$$

Dynamic scaling:

$$\alpha = \beta z$$

β growth exponent

z dynamic exponent

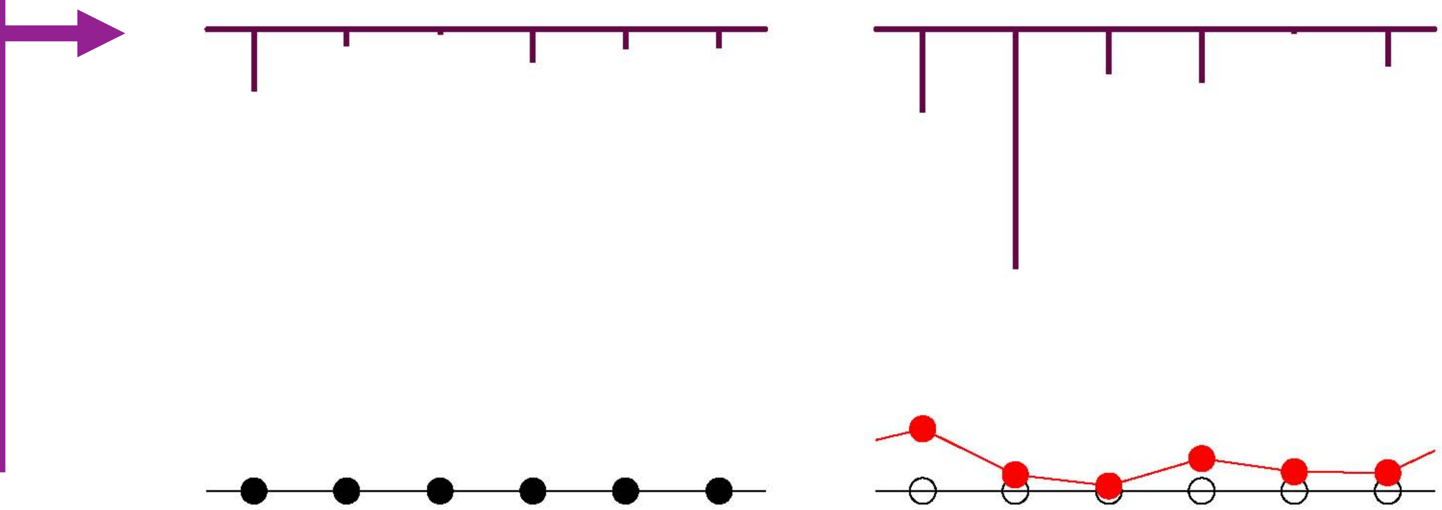
α roughness exponent

Non-equilibrium surface growth model: **PDES model**

$-ln(r)$

$0<r<1$

Replenish when needed



Start with flat interface *(in d dimensions)*

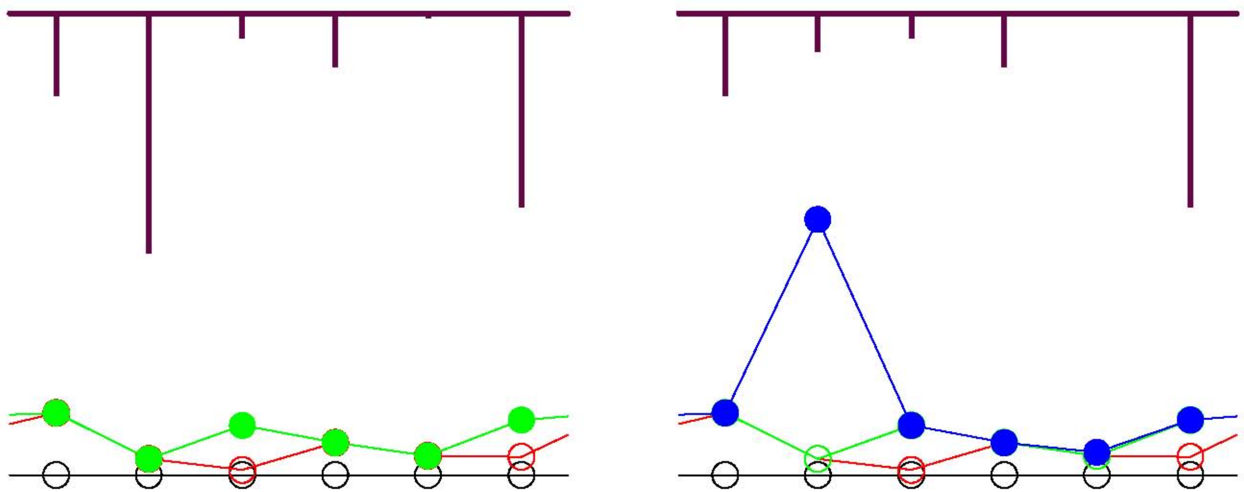
In first step, all 'drops' fall

PDES *model*

-ln(r)

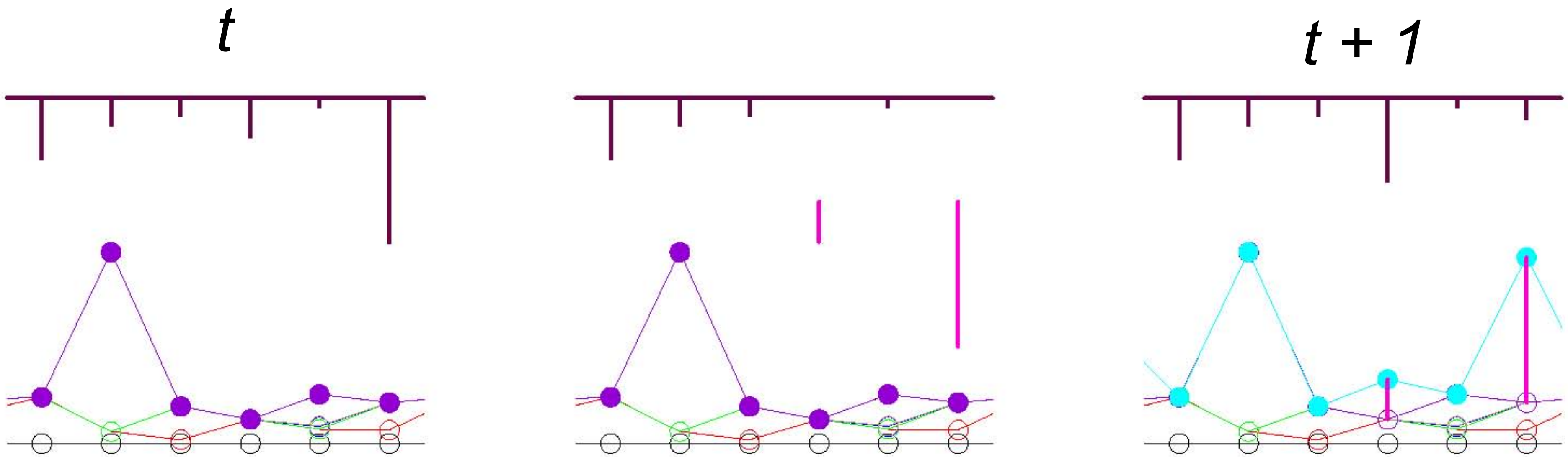
0<r<1

Replenish when needed



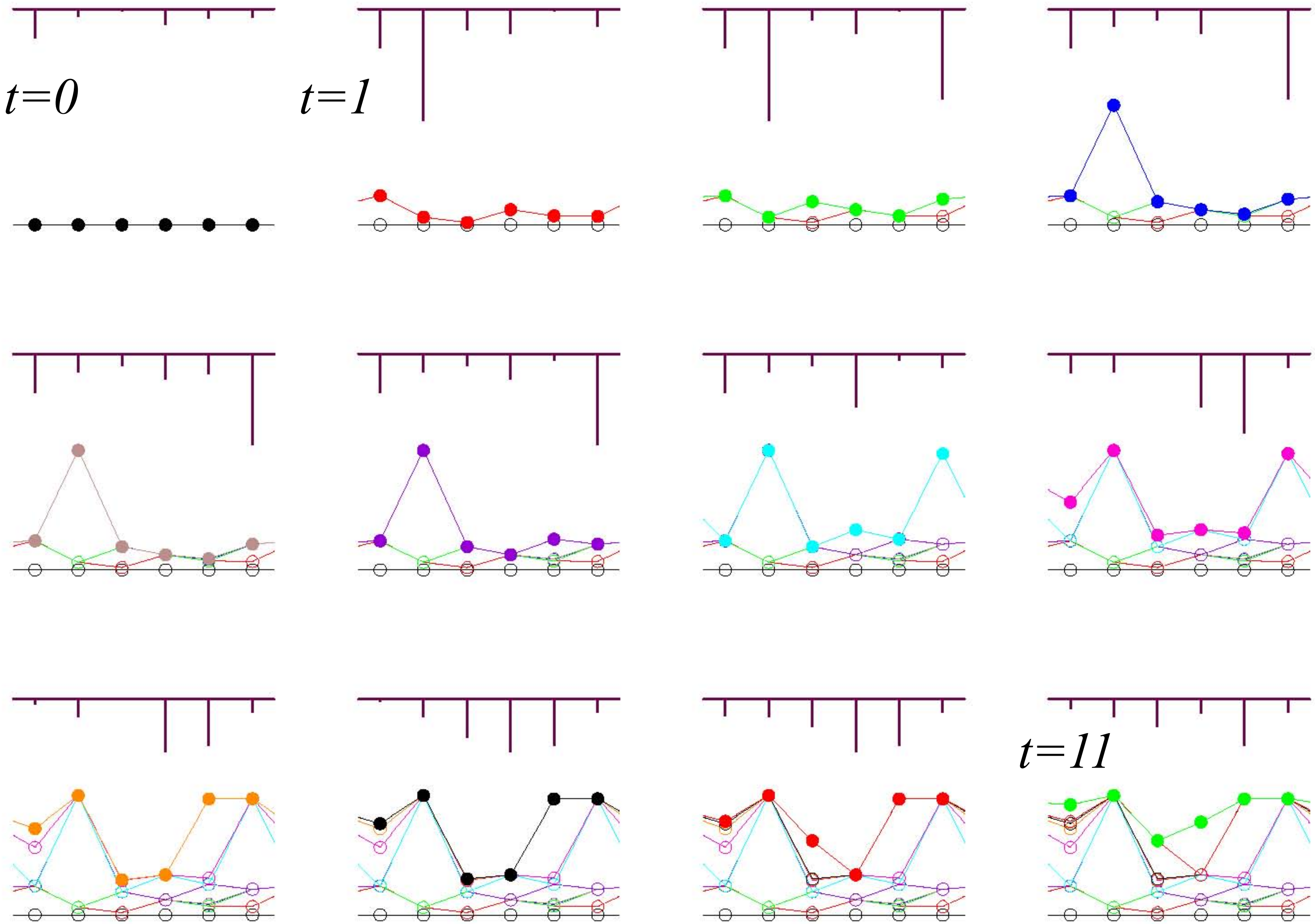
For each step, all 'drops' fall ***ONLY*** if the surface underneath is at a local minimum

PDES *model*



Note: at each step *t* all 'drops' fall ***at the same time***

PDES *model*



Discrete Event Simulations

- DES (Discrete Event Simulations)
  * State changes are discontinuous
  * Times of state changes are random

PDES

Parallel Discrete Event Simulations

PDES Technology Implications

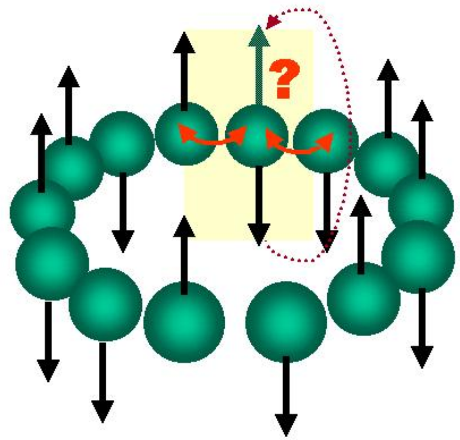
- **All today's largest computers are massively parallel computers**
- **Must make good use of parallelization in programs for efficiency**
- **Parallel Discrete Event Simulations (PDES)**
    - o **Used in military simulations and training ('what-if' scenarios)**
    - o **Used in homeland security simulations and training**
    - o **Used in modeling of factory deliveries**
    - o **Used in modeling temporal drug concentrations in patient models**
    - o **Used in simulating materials and materials failure**
    - o **Used in modeling switching in cellular and wireless networks**
    - o **Used in ecological modeling**
    - o **Used in modeling epidemiological models**
    - o **Used in electric power grid simulations**

Information-Driven Systems

Example:
Dynamic Monte Carlo of Ising spins with nearest-neighbor interactions

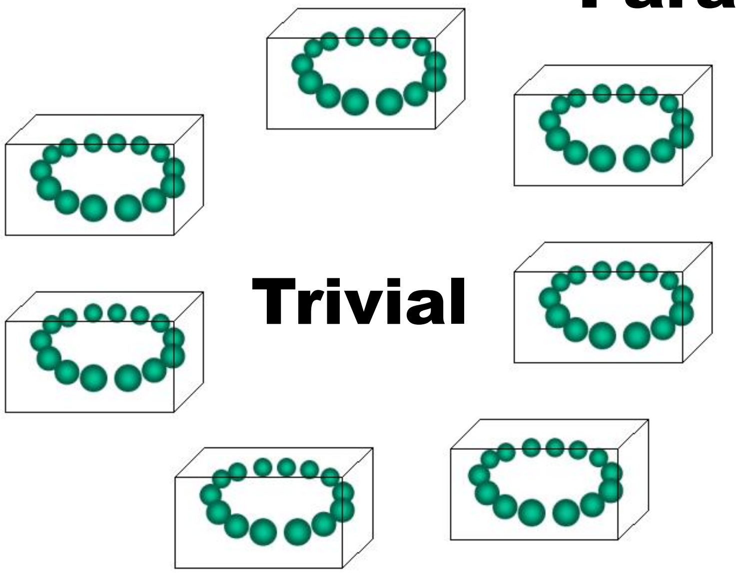
Randomly pick a spin

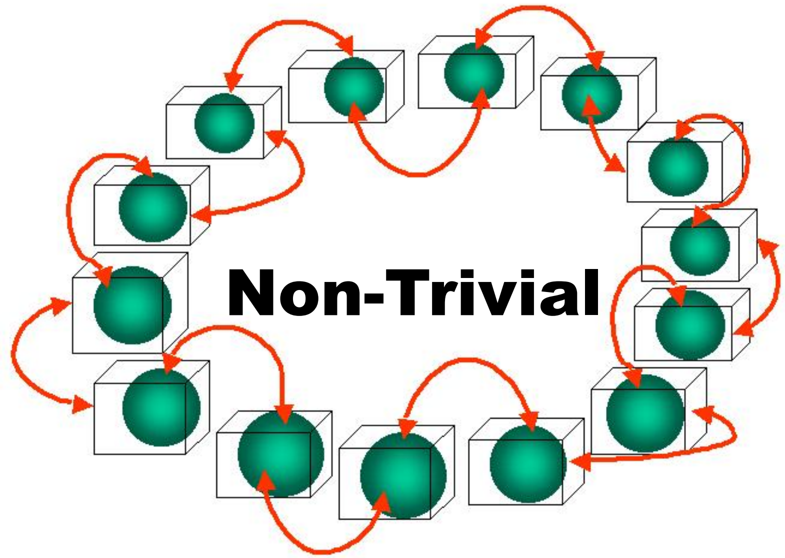
Decide if spin will be flipped



Dynamic Monte Carlo simulations

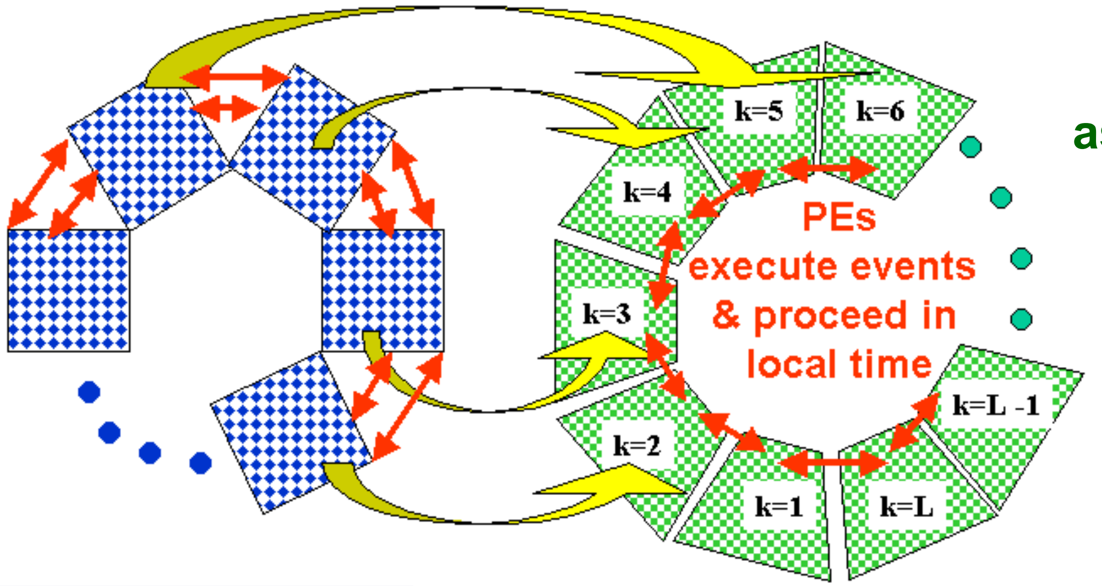
Parallelization





Physical processes and logical processes

asynchronous nature of physical dynamics



asynchronous system of logical processes

Physical System		Computing System
spatially extended system of *NL* spins, arranged on a lattice	→	*L* PEs: each carries *N* lattice sites, N_b of which are border sites
Physical Events/Processes random spin flipping	→	**Logical Events/Processes** each PE manages the state of the assigned subsystem.

discrete event: the spin flip → discrete event: the state update

Parallel discrete-event simulation

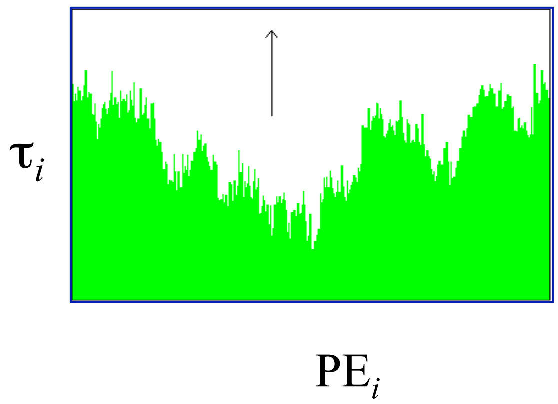
for spatially decomposable asynchronous cellular automata

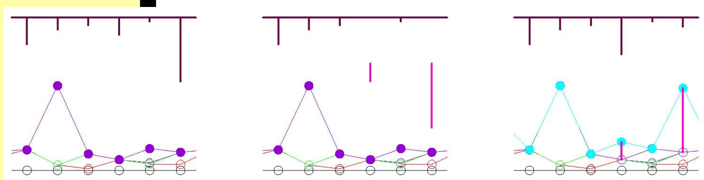
- Spatial decomposition on lattice/grid
  (for systems with short-range interactions
  only local synchronization between subsystems)
- Changes/updates: independent Poisson arrivals

❖ Each subsystem/block of sites, carried by a processing element (PE) must must have its own local simulated time, $\{\tau_i\}$ ("virtual time")

❖ Synchronization scheme

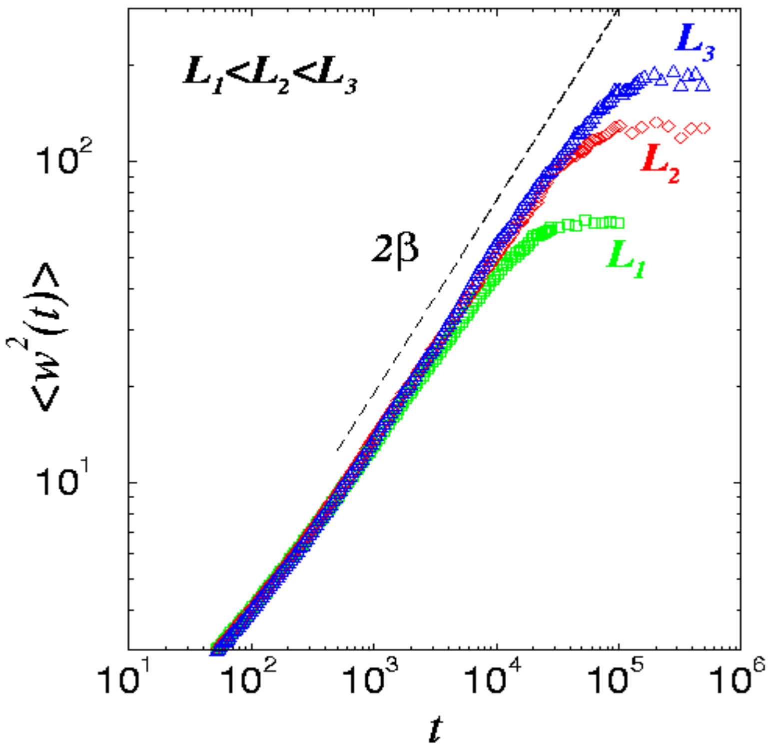
❖ PEs must concurrently advance their own Poisson streams, without violating causality



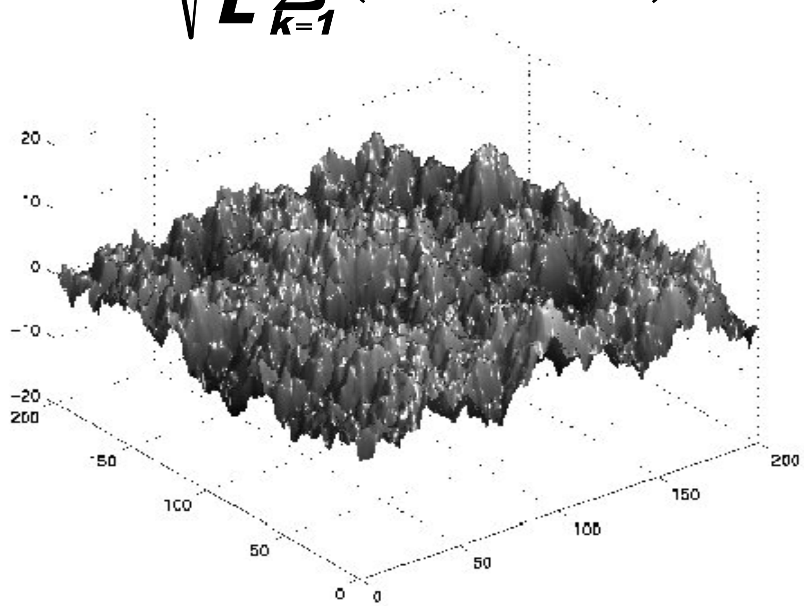


This *is* the PDES model

Non-equilibrium surface growth



$$\boldsymbol{w}(\boldsymbol{t}) = \sqrt{\frac{1}{L}\sum_{k=1}^{L}\left(\tau_k(\boldsymbol{t}) - \bar{\tau}(\boldsymbol{t})\right)^2}$$



$$\langle w^2(t)\rangle \sim \begin{cases} t^{2\beta}, & if \quad t << t_\times \\ L^{2\alpha}, & if \quad t >> t_\times \end{cases}$$

Dynamic scaling:

$$\alpha = \beta z$$

β growth exponent

$\boldsymbol{z}$ dynamic exponent

α roughness exponent

Coarse graining for the stochastic time surface evolution

Korniss, Toroczkai, Novotny, Rikvold, PRL '00

$$\tau_i(t+1) = \tau_i(t) + \eta_i(t)\Theta[\tau_{i-1}(t) - \tau_i(t)]\Theta[\tau_{i+1}(t) - \tau_i(t)]$$

- $\Theta(\ldots)$ is the Heaviside step-function
- $\eta_i(t)$ iid exponential random numbers

$\vdots$

$$\frac{\partial h(x,t)}{\partial t} = \nu \frac{\partial^2 h(x,t)}{\partial x^2} + \lambda \left[\frac{\partial h(x,t)}{\partial x}\right]^2 + D_{kpz}\, \eta(x,t)$$

Kardar-Parisi-Zhang equation

$$P[\tau(x)] \propto \exp\left[-\frac{1}{2D}\int dx \left(\frac{\partial \tau}{\partial x}\right)^2\right]$$

Steady state (d=1):
Edwards-Wilkinson Hamiltonian

❖Random-walk profile: short-range correlated local slopes

"Simulating the simulations"

❖Universality/roughness (*d*=1)

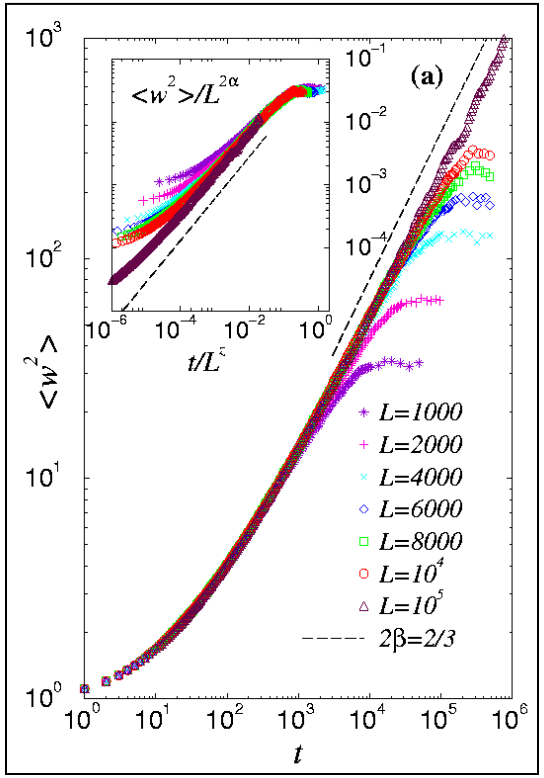
$$\langle w^2(t)\rangle_L \sim \begin{cases} t^{2\beta}, & \text{if} \quad t << t_\times \\ L^{2\alpha}, & \text{if} \quad t >> t_\times \end{cases}, \quad t_\times \sim L^z, \quad z = \alpha/\beta$$

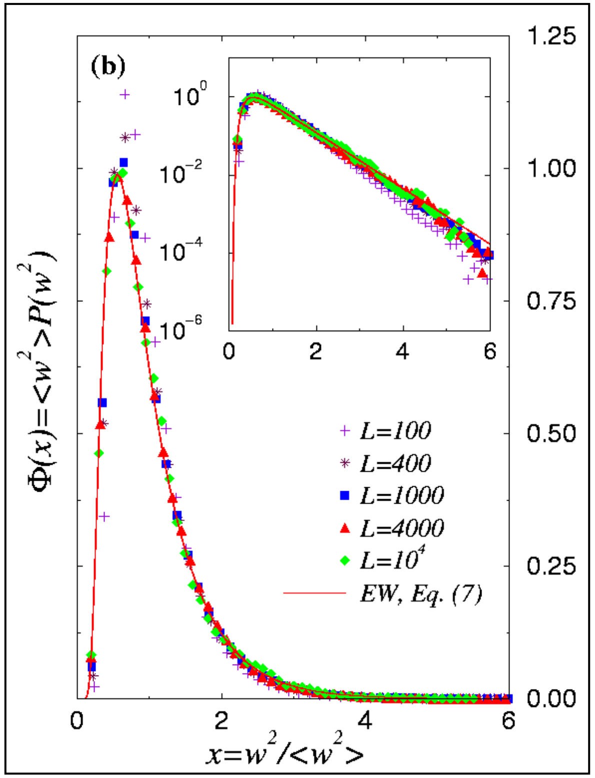
Foltin et.al., '94

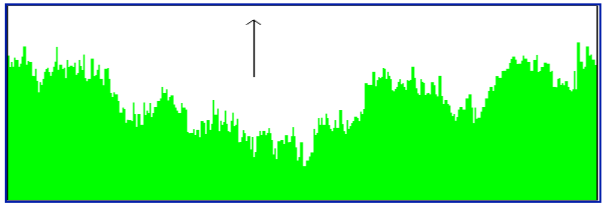
$$\beta \approx 0.33, \quad \alpha \approx 0.5$$

$$P(w^2) = \langle w^2\rangle^{-1}\Phi\left(w^2/\langle w^2\rangle\right)$$

exact KPZ:
$\beta=1/3$
$\alpha=1/2$



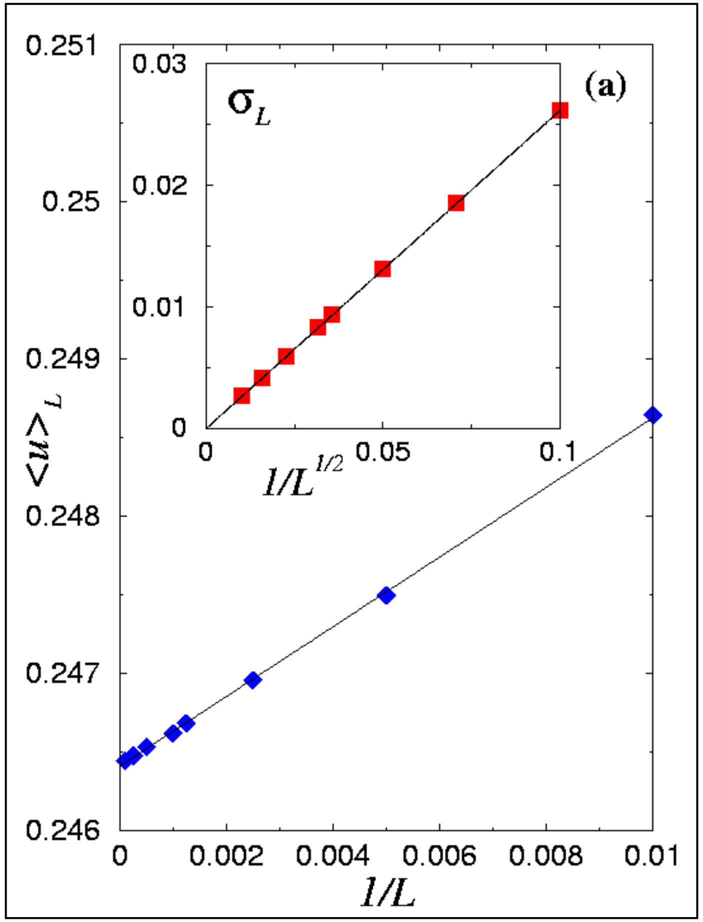




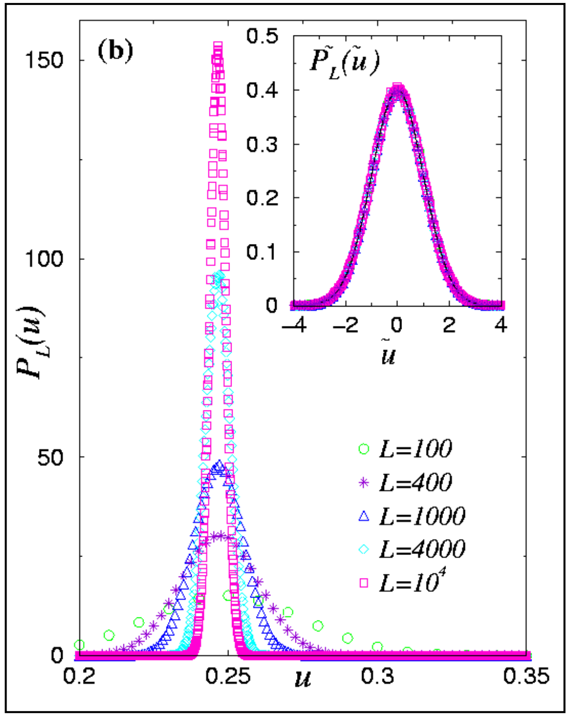
❖Utilization/efficiency

Finite-size effects for the density of local minima/average growth rate (steady state):

$$\langle u \rangle = \langle u \rangle_\infty + \text{const}/\text{N}_{\text{PE}}$$



$$\sigma_L = \sqrt{\langle u^2 \rangle_L - \langle u \rangle_L^2} \sim 1/L^{1/2}$$



$d=1$

$$\langle u \rangle_\infty \approx 0.2464$$

Implications for scalability

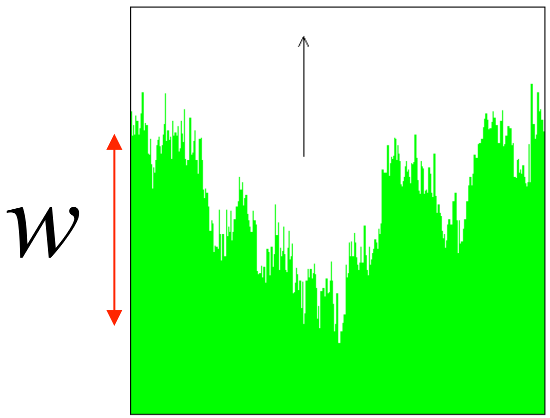
Virtual Time Horizon belongs to KPZ universality class

GREAT News ----------- **Bad** News

❖Simulation phase: scalable $\langle u \rangle = \langle u \rangle_{\infty} + \mathrm{const}/N_{PE}$

$\langle u \rangle_{\infty}$ asymptotic average rate of progress of the simulation (utilization) is non-zero

❖Measurement (data management) phase: not scalable



$$\langle w^2(t) \rangle_L \sim \begin{cases} t^{2\beta}, & if \quad t << t_{\times} \\ L^{2\alpha}, & if \quad t >> t_{\times} \end{cases}$$

Rough Times!

Improve efficiency

Mixing

KPZ + RD

$$\frac{\partial h(x,t)}{\partial t} = \nu \frac{\partial^2 h(x,t)}{\partial x^2} + \lambda \left[\frac{\partial h(x,t)}{\partial x} \right]^2 + D_{kpz}\, \eta(x,t)$$

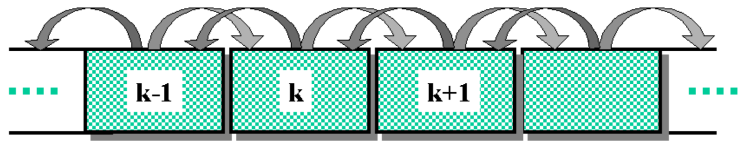
$$+$$

$$\frac{\partial h(x,t)}{\partial t} = D_{rd}\, \eta(x,t)$$

Simulation model for conservative PDES

Time-step *t* : index of the simultaneous update attempt
Updates at *t* : independent Poisson-random processes

If update at *t* : $h_k(t+1) = h_k(t) + \eta_k(t)$



Update rule

N=1

$h_k(t) \leq min\{h_{k-1}(t), h_{k+1}(t)\}$

N=2

choose a neighbor

$h_k(t) \leq h_{nn}(t)$

N>2

choose a lattice site

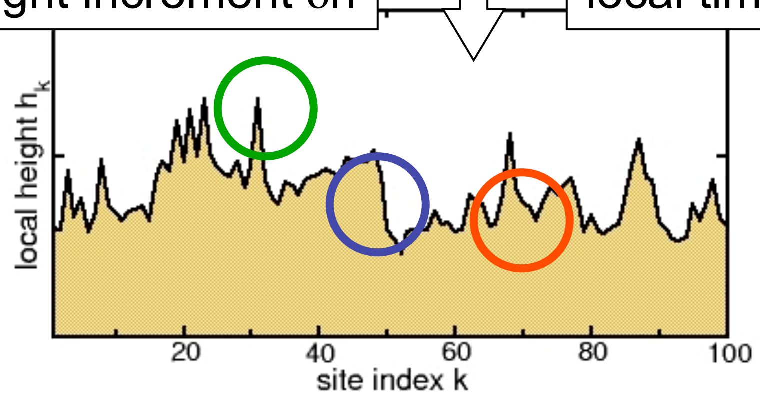
interior

border

$h_k(t) \leq h_{nn}(t)$

deposition at t
local height increment δh

update at t
local time increment δh

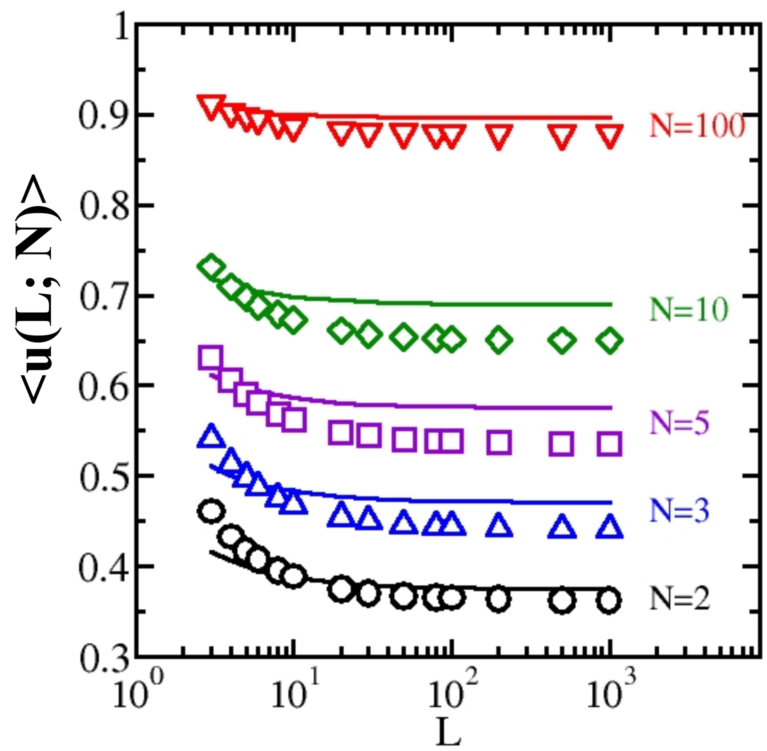


Virtual Time Horizon (VTH)

Properties of the algorithm are encoded in the VTH

Diagnostics: utilization of the parallel processing environment

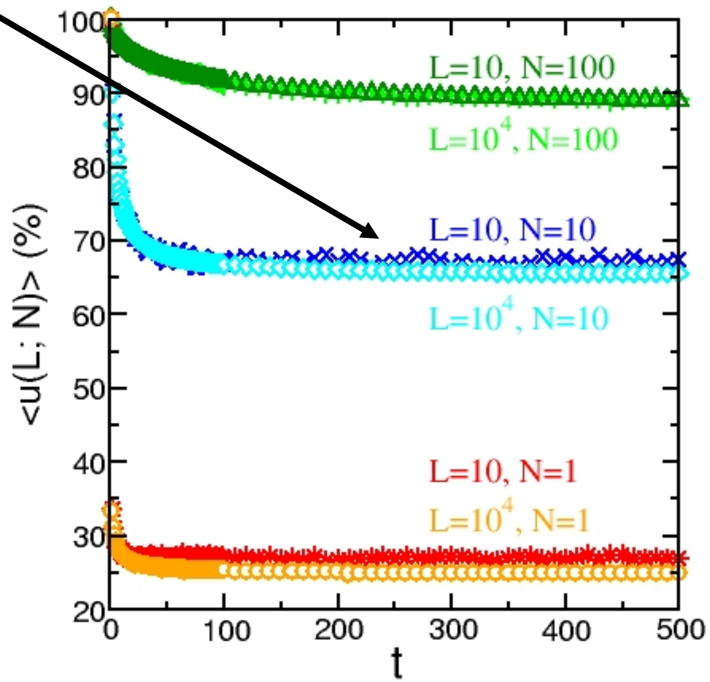
Steady-state simulations



$$\langle u\,(L>2;\,N>1)\rangle =$$

$$= \left(1-\frac{1}{\sqrt{2N}}\right)\left(1-\frac{1}{2\sqrt{2N}}\frac{L-1}{L}\right)$$

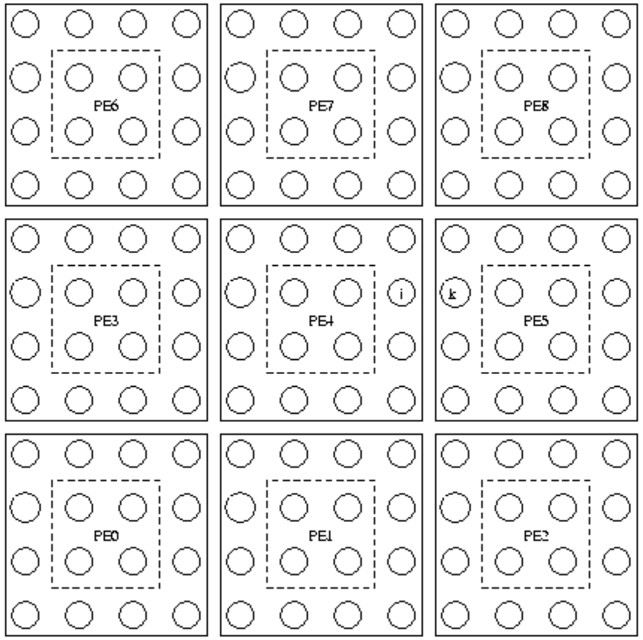
$v(t) = \langle u(t)\rangle\, \mu_P$

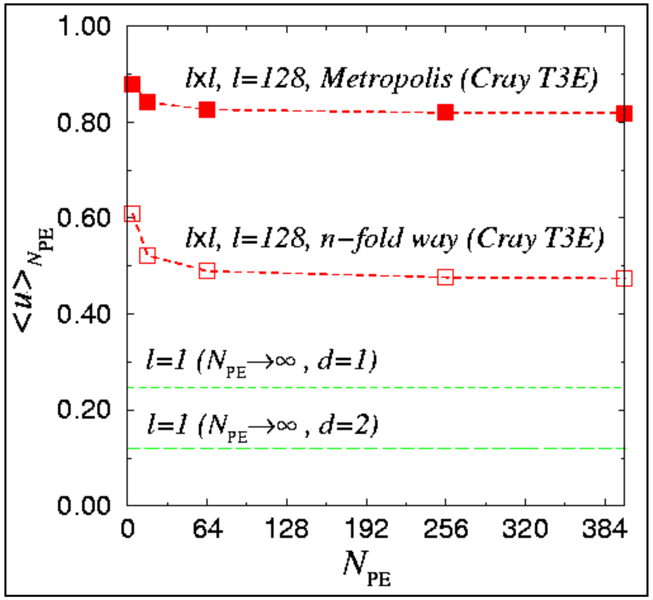


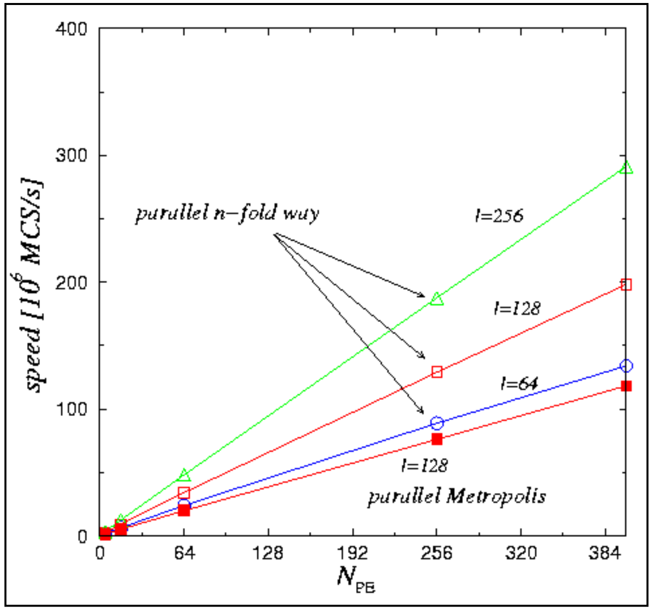
PRB **69**, 075407 (2004)

Actual implementation

Dynamics of a thin magnetic film







1. Local time incremented
2. Randomly chosen site
3. If chosen site is on the boundary, PE must wait until $\tau \leq \min\{\tau_{nn}\}$

l > 1 → Mixing RD+KPZ

Implications for scalability

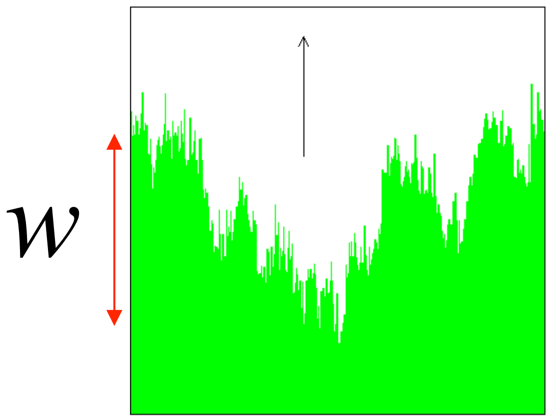
Virtual Time Horizon belongs to KPZ universality class

GREAT News ----------- **Bad** News

- Simulation phase: scalable $\langle u \rangle = \langle u \rangle_\infty + \text{const}/N_{PE}$

$\langle u \rangle_\infty$ asymptotic average rate of progress of the simulation (utilization) is non-zero

- Measurement (data management) phase: not scalable



$$\langle w^2(t) \rangle_L \sim \begin{cases} t^{2\beta}, & if \quad t << t_\times \\ L^{2\alpha}, & if \quad t >> t_\times \end{cases}$$

Rough Times!

How to make the measurement phase scalable as well?

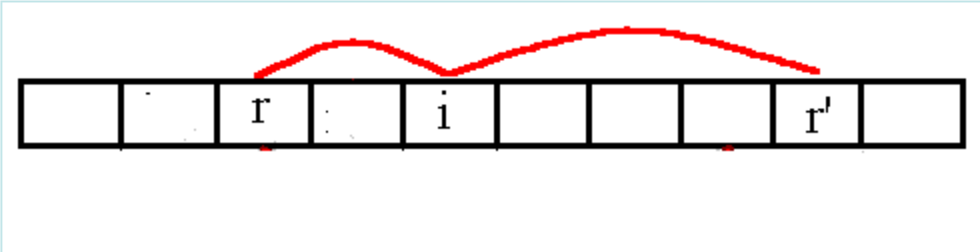
❖Controlling the width

Greenberg et.al., '96

"Mean-field" like approximation to model the evolution of the time horizon (K-random interaction)

K=2: each PE randomly chooses two others, *r* and *r'*

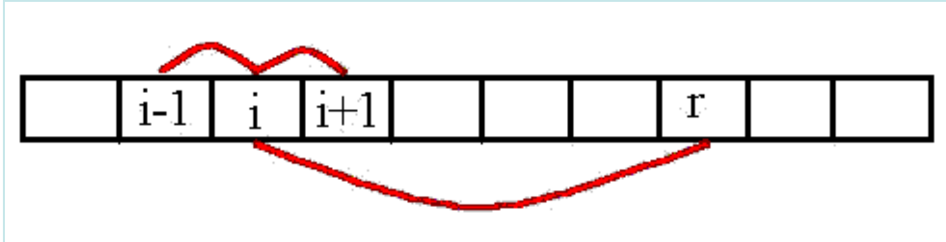
$$\tau_i \leq \min\{\tau_r, \tau_{r'}\}$$



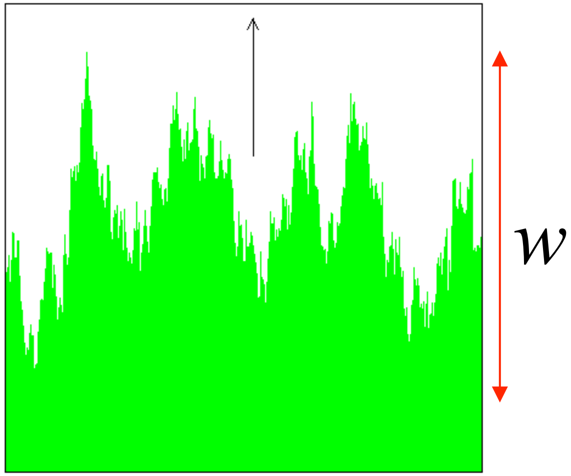
❖Width is finite in this mean-field model when $L \to \infty$

❖$\langle u \rangle_L \sim 1/(K+1)$ is nonzero

Annealed (or quenched) random connections

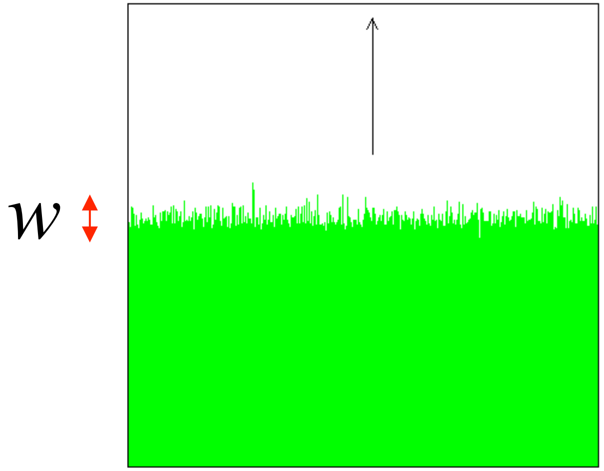


$\tau_i \le \min\{\tau_{\rm nn}\}$



KPZ surface: $w \sim L^{\alpha}$

$\tau_i \le \min\{\tau_{\rm nn}, \tau_r\}$

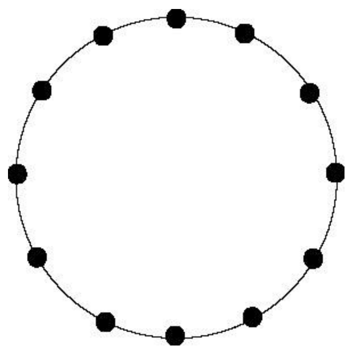


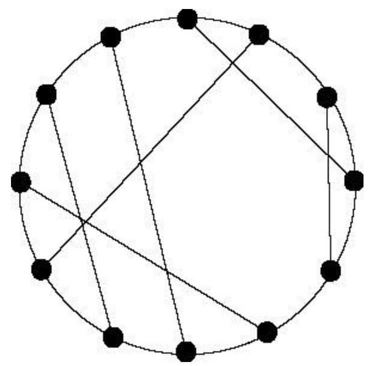
$w = const. + O(L^{-1})$

$$\partial_t \tau = \frac{\partial^2 \tau}{\partial x^2} - \lambda \left(\frac{\partial \tau}{\partial x} \right)^2 + \ldots + noise$$

$$\partial_t \tau = -\gamma(\tau(x,t) - \bar{\tau}(t)) + \frac{\partial^2 \tau}{\partial x^2} + \ldots + noise$$

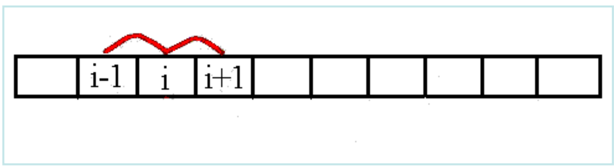
Slopes are still short-range correlated: non-zero $\langle u \rangle$

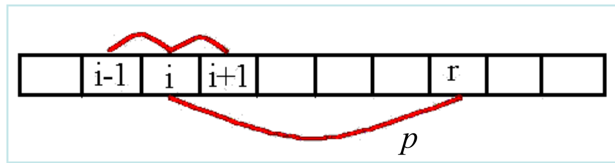


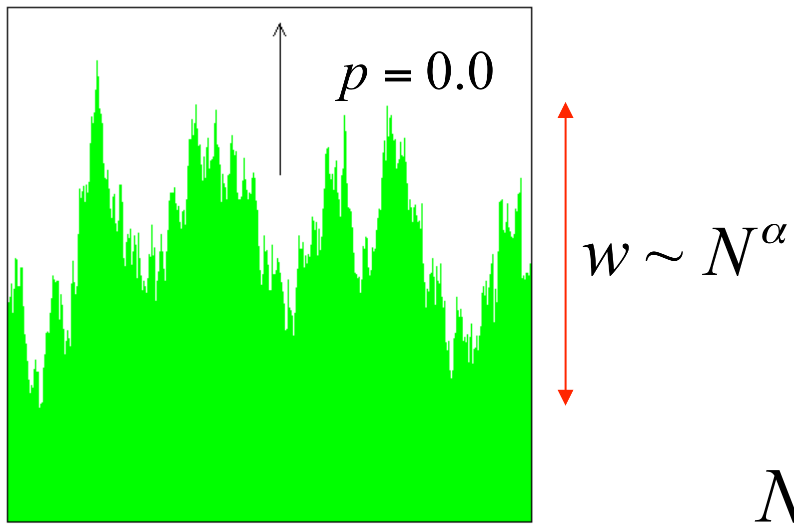


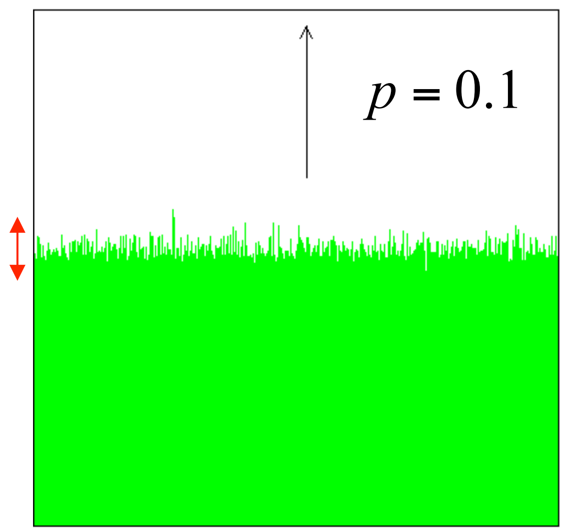
regular lattice (ring) topology
("*p=0*")

small-world-like connections:
used with probability *p>0*









$$N = 10^4$$

$$w \underset{N\to\infty}{\sim} const.$$

Steady-state “height” structure factors

$$S(k) \propto \langle \tau(k)\tau(-k) \rangle$$

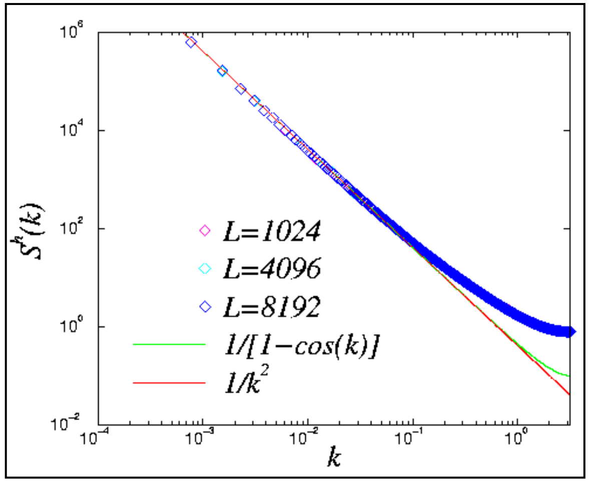
only short-range connections (KPZ)

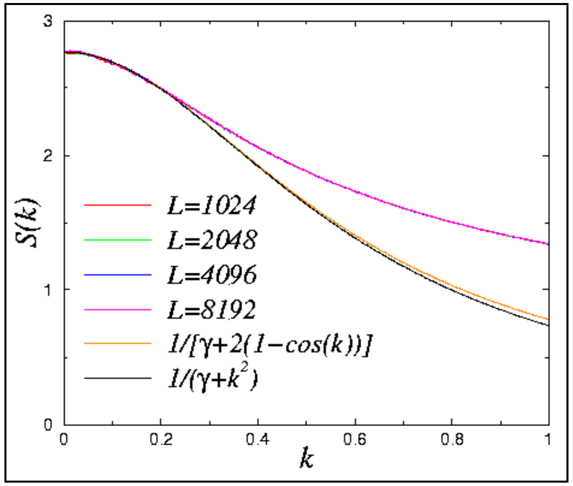
$$\partial_t \tau = \frac{\partial^2 \tau}{\partial x^2} - \lambda \left(\frac{\partial \tau}{\partial x} \right)^2 + noise$$

+ random connections (relaxation)

$$\partial_t \tau = -\gamma(\tau(x,t) - \overline{\tau}(t)) + \frac{\partial^2 \tau}{\partial x^2} + noise$$

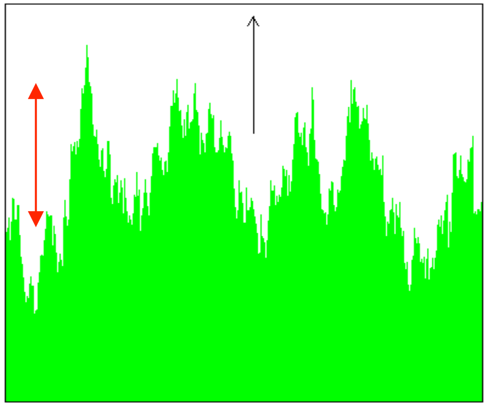
(d=1) $$S(k) \sim \frac{1}{k^2}$$ $$S(k) \sim \frac{1}{\gamma + k^2}$$

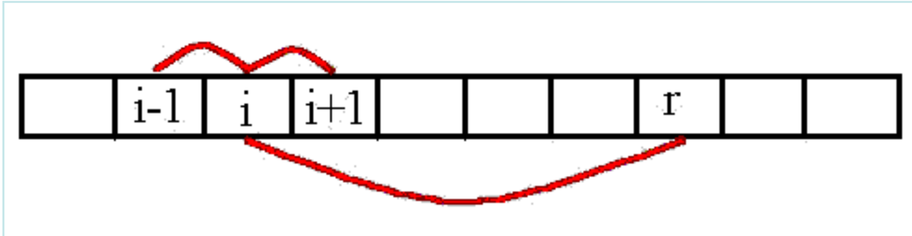




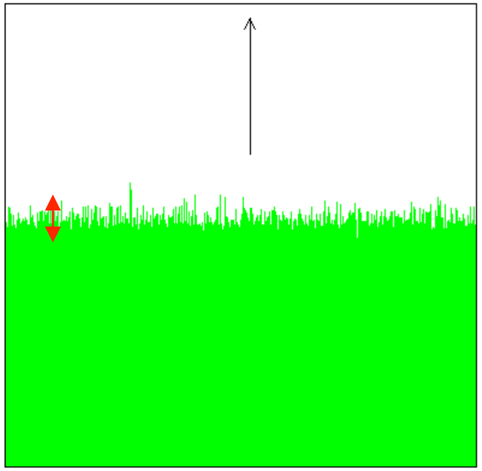
Quenched random (Small World) connections

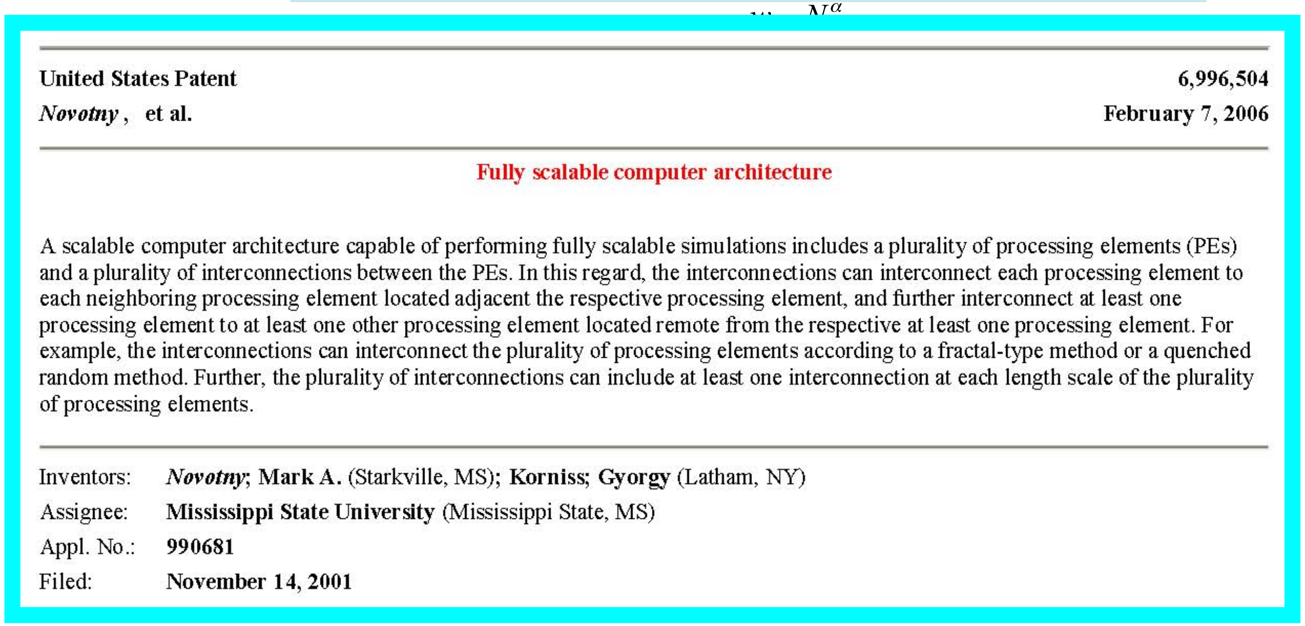
w





w





United States Patent **6,996,504**

***Novotny*, et al.** **February 7, 2006**

Fully scalable computer architecture

A scalable computer architecture capable of performing fully scalable simulations includes a plurality of processing elements (PEs) and a plurality of interconnections between the PEs. In this regard, the interconnections can interconnect each processing element to each neighboring processing element located adjacent the respective processing element, and further interconnect at least one processing element to at least one other processing element located remote from the respective at least one processing element. For example, the interconnections can interconnect the plurality of processing elements according to a fractal-type method or a quenched random method. Further, the plurality of interconnections can include at least one interconnection at each length scale of the plurality of processing elements.

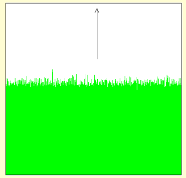
Inventors: ***Novotny*; Mark A.** (Starkville, MS); **Korniss; Gyorgy** (Latham, NY)
Assignee: **Mississippi State University** (Mississippi State, MS)
Appl. No.: **990681**
Filed: **November 14, 2001**

$$w = const. + O(L^{-1}) \qquad \tau_i \le \min\{\tau_{\mathrm{nn}}, \tau_r\}$$

Slopes are still short-range correlated: non-zero $\langle u \rangle$

PDES Summary and outlook

- Simple surface-growth model very useful



- The tools and methods of non-equilibrium statistical physics (coarse-graining, finite-size scaling, universality, etc.) can be applied to scalability modeling and algorithm engineering
- Conservative schemes can be made perfectly scalable (ALL short-ranged PDES)
  - Computational phase always scalable (KPZ universality)
  - Communication phase scalable with small-world network

Fully Scalable Computer Architectures

Main Paper

REPORTS

Suppressing Roughness of Virtual Times in Parallel Discrete-Event Simulations

G. Korniss,[1*] M. A. Novotny,[2] H. Guclu,[1] Z. Toroczkai,[3] P. A. Rikvold[4]

In a parallel discrete-event simulation (PDES) scheme, tasks are distributed among processing elements (PEs) whose progress is controlled by a synchronization scheme. For lattice systems with short-range interactions, the progress of the conservative PDES scheme is governed by the Kardar-Parisi-Zhang equation from the theory of nonequilibrium surface growth. Although the simulated (virtual) times of the PEs progress at a nonzero rate, their standard deviation (spread) diverges with the number of PEs, hindering efficient data collection. We show that weak random interactions among the PEs can make this spread nondivergent. The PEs then progress at a nonzero, near-uniform rate without requiring global synchronizations.

Simulating large systems often leaves the programmer with only one option: parallel distributed simulations where parts of the system are allocated and simulated on different processing elements (PEs). A large class of interacting systems, including financial market models, epidemic models, dynamics of magnetic systems, and queuing networks, can be described by a set of local state variables assuming a finite number of possible values. As the system evolves in time, the values of the local state variables change at discrete instants, synchronously or asynchronously depending on the dynamics of the system. Parallel simulation for the former is straightforward (at least conceptually). For the latter—that is, for asynchronous or nonparallel dynamics—one must use some kind of synchronization to ensure causality. The instantaneous changes in the local configuration are also called discrete events, hence the term parallel discrete-event simulation (PDES) (*1–3*). Examples of PDES applications include dynamic channel allocation in cell phone communication networks (*3, 4*), models of the spread of diseases (*5*), battlefield simulations (*6*), and dynamic phenomena in highly anisotropic magnetic systems (*7, 8*). Here the discrete events are call arrivals, infections, troop movements, and changes of the orientation of the local magnetic moments, respectively. As the number of PEs on parallel architectures increases to tens of thousands, fundamental questions of the scalability of the underlying algorithms must be addressed. Here, we show a way to construct fully scalable parallel simulations for systems with asynchronous dynamics and short-range interactions. Understanding the effects of the microscopic dynamics (corresponding to the algorithmic synchronization rules) on the global properties of the simulation scheme brings us to the solution. Recently, a similar connection has been made (*9*) between rollback-based PDES schemes (*10*) and self-organized criticality (*11*).

The two basic ingredients of PDES are the set of local simulated times, often referred to as virtual times (*10*), and a synchronization scheme (*1*). For the PDES scheme to be scalable (*12*), two criteria must be met: (i) The virtual time horizon should progress on average at a nonzero rate, and (ii) the typical spread of the time horizon should be bounded as the number of PEs N_{PE} goes to infinity. The first criterion ensures a nonzero progress rate in the limit of large N_{PE}. It is, however, not sufficient if data are to be collected. Different PEs have progressed to different local simulated times with a possibly large spread among them, making measurement a complex task. Frequent global synchronizations can get costly for large N_{PE}, whereas temporarily storing a large amount of data as a result of the large virtual time spread is limited by the available memory. Therefore, criterion (ii) is crucial for the measurement part of the algorithm to be scalable. Here, we introduce a PDES scheme in which the PEs make nonzero and close-to-uniform progress without global intervention.

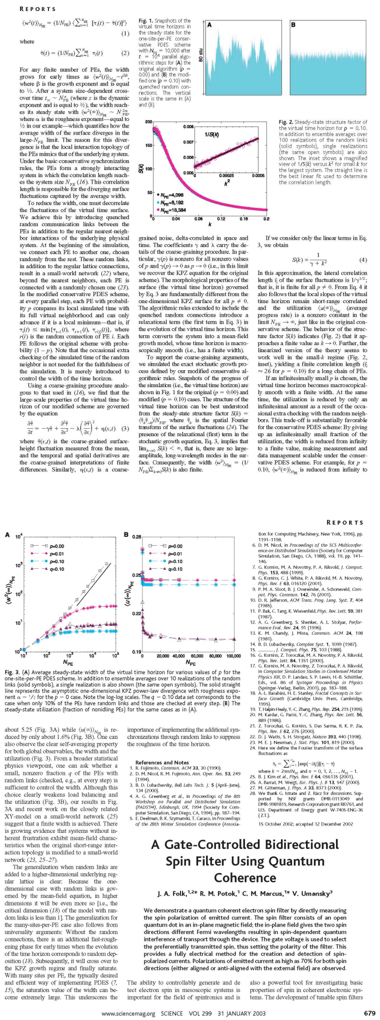
In conservative PDES schemes (*13–15*), which we focus on, an update is performed by a particular PE only if the resulting change in the local configuration of the simulated system is guaranteed not to violate causality. Otherwise, the PE idles. The efficiency of the scheme depends on the fraction of nonidling PEs. It was shown (*16, 17*) that the virtual time horizon exhibits kinetic roughening (*18, 19*) for the basic conservative scheme applied to systems with short-range interactions on regular lattices. In particular, the evolution of the virtual time horizon is governed by the Kardar-Parisi-Zhang (KPZ) equation (*20*), which plays a central role in nonequilibrium surface growth (*18, 19*). The above finding has two major implications for the asymptotic scalability of the basic conservative PDES scheme (*16, 21*): Criterion (i) for the scalability is satisfied because the average progress rate of the virtual time horizon approaches a nonzero value in the limit $N_{PE} \to \infty$. Criterion (ii), however, is violated because the virtual time horizon becomes macroscopically rough.

For illustration, we consider a general one-dimensional system with nearest-neighbor interactions, in which the discrete events exhibit Poisson asynchrony. In the one-site-per-PE scenario, each site has its own local simulated time, constituting the virtual time horizon $\{\tau_i(t)\}_{i=1}^{N_{PE}}$, where t is the discrete number of parallel steps executed by all PEs (which is proportional to the wall-clock time). According to the basic conservative synchronization scheme (*14, 15*), at each parallel step t, only those PEs for which the local simulated time is not greater than the local simulated times of their neighbors can increment their local time by an exponentially distributed random amount. [Without loss of generality, we assume that the mean of the local time increment is 1 in simulated time units (stu).] Thus, if $\tau_i(t) \leq \min\{\tau_{i-1}(t), \tau_{i+1}(t)\}$, PE i can update the configuration of the underlying site it carries and determine the time of the next event. Otherwise, it idles. Despite its simplicity, this rule preserves unaltered the asynchronous causal dynamics of the underlying system (*14, 15*).

The progress rate of the simulation $\langle u(t)\rangle_{N_{PE}}$ (the density of local minima of the virtual time horizon) approaches a nonzero constant in the asymptotic long-time, large-N_{PE} limit (*16, 21*). The average width of the virtual time horizon, however, diverges as $N_{PE} \to \infty$ (*16, 17*). Specifically, the average width is defined as

[1]Department of Physics, Applied Physics, and Astronomy, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA. [2]Department of Physics and Astronomy and ERC Center for Computational Science, Mississippi State University, Post Office Box 5167, Mississippi State, MS 39762, USA. [3]Complex Systems Group, Theoretical Division, Los Alamos National Laboratory, Mail Stop B-213, Los Alamos, NM 87545, USA. [4]Department of Physics, Center for Materials Research and Technology, and School of Computational Science and Information Technology, Florida State University, Tallahassee, FL 32306, USA.

*To whom correspondence should be addressed. E-mail: korniss@rpi.edu



US Patent

6,996,504

Issued Feb. 7, 2006

Novotny & Korniss

Large Paper

Guclu *et al*, PRE 2006

Shows *ALL* parallel discrete event simulations that are short-ranged can be made to be perfectly scalable on the correct computer architecture.

Discussion and Provocations

- Neither software nor hardware nor algorithms alone will lead to (non-trivial) perfect scalability
- Without use of statistical mechanics, parallel computing will never be efficient/scalable

- Similar ideas apply to (non-trivial) cloud computing
- Similar ideas for sensor networks
- Similar ideas for databases and searches
- Similar ideas for fault-tolerant computing

- Similar ideas can be used to design new materials and devices with novel properties

Novotny
Selected References

1. **Synchronization in small-world-connected computer networks,** Guclu *et al*, PRE **73**, 066115 (2006).
2. **Universal scaling in mixing correlated growth with randomness** Kolakowska *et al*, PRE **73**, 011603 (2006).
3. **Desynchronization and speedup in an asynchronous conservative parallel update protocol**, Kolakowska & Novotny, Ch. 6 in "Artificial Intelligence and Computer Science" (Nova Science 2005).
4. **Evolution of Time Horizons in Parallel and Grid Simulations,** L.N. Shchur and M.A. Novotny, PRE **70**, 026703 (2004).
5. **Roughening of the interfaces in (1+1) dimensional two-component surface growth with an admixture of random deposition**, Kolakowska *et al*, PRE **70**, 051602 (2004).
6. **Discrete-event analytic technique for surface growth problems**, Kolakowska and Novotny, PRB **69**, 075407 (2004).
7. **Suppresssing roughness of virtual times in parallel discrete-event simulations,** Korniss, Novotny, Guclu, Toroczkai, Rikvold,SCIENCE **299,** 677 (2003).
8. **Update statistics in conservative PDES**, Kolakowska *et al*, PRE **68**, 046705 (2003).
9. **Algorithmic scalability in globally constrained conservative PDES** Kolakowska *et al*, PRE **67**, 046703 (2003).
10. **Algorithms for faster and larger dynamic Metropolis simulations** Novotny *et al*, AIP Conference proceedings, 2003.
11. **Statistical Properties of the simulated time horizon in conservative PDES** Korniss *et al*, ACM Proceedings, 2002.