Virtualization-aware Application Framework for High-end Classical-quantum Atomistic Simulations of Nanosystems

Aiichiro Nakano

Collaboratory for Advanced Computing & Simulations
Department of Computer Science, Department of Physics & Astronomy,
Department of Materials Science & Engineering
University of Southern California

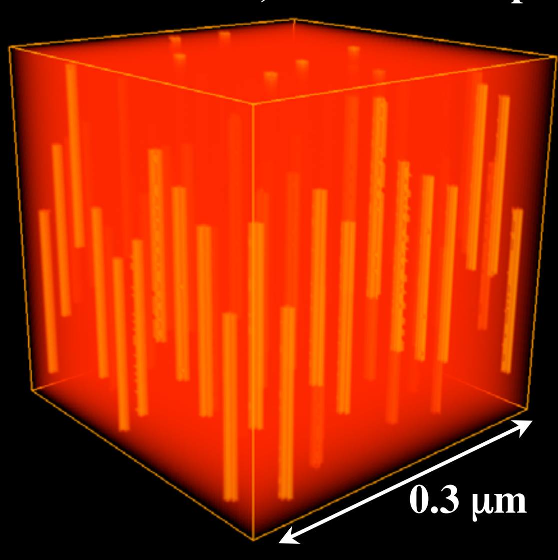
Email: anakano@usc.edu

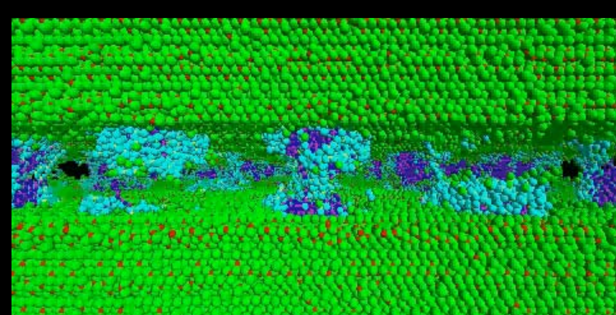
Collaborators:

**Rajiv K. Kalia, Ashish Sharma, Priya Vashishta (*CACS*),
Hiroshi Iyetomi (*Niigata*), Hideaki Kikuchi (*Louisiana*),
Shuji Ogata (*Nagoya Inst. Tech.*), Fuyuki Shimojo (*Kumamoto*),
Kenji Tsuruta (*Okayama*)**

High End Computing Is Bringing Atomistic Simulation To Macroscopic

1.5 billion atom molecular dynamics simulation on 1,024 IBM SP3 processors at NAVO-MSRC



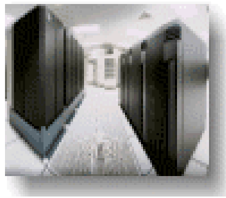


Color code: Si_3N_4; SiC; SiO_2

Computing beyond Teraflop: Grid of distributed supercomputers?

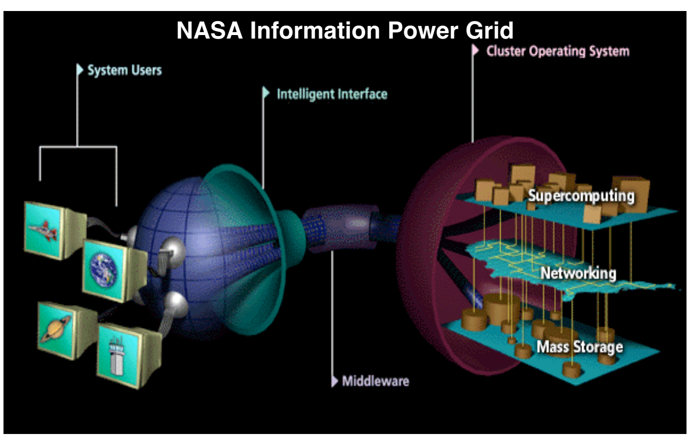
Parallel Benchmark Platforms

DoD-NAVO
IBM SP4



1	**NEC** Earth-Simulator/ 5120	**35860.00** 40960.00	Earth Simulator Center Japan/2002
2	**Hewlett-Packard** ASCI Q - AlphaServer SC ES45/ 1.25 GHz/ 4096	**7727.00** 10240.00	Los Alamos National Laboratory USA/2002
3	**Hewlett-Packard** ASCI Q - AlphaServer SC ES45/ 1.25 GHz/ 4096	**7727.00** 10240.00	Los Alamos National Laboratory USA/2002
4	**IBM** ASCI White, SP Power3 375 MHz/ 8192	**7226.00** 12288.00	Lawrence Livermore National Laboratory USA/2000
5	**Linux NetworX** MCR Linux Cluster Xeon 2.4 GHz - Quadrics/ 2304	**5694.00** 11060.00	Lawrence Livermore National Laboratory USA/2002
6	**Hewlett-Packard** AlphaServer SC ES45/1 GHz/ 3016	**4463.00** 6032.00	Pittsburgh Supercomputing Center USA/2001
7	**Hewlett-Packard** AlphaServer SC ES45/1 GHz/ 2560	**3980.00** 5120.00	Commissariat a l'Energie Atomique (CEA) France/2001
8	**HPTi** Aspen Systems, Dual Xeon 2.2 GHz - Myrinet2000/ 1536	**3337.00** 6758.00	Forecast Systems Laboratory - NOAA USA/2002
9	**IBM** pSeries 690 Turbo 1.3GHz/ 1280	**3241.00** 6656.00	HPCx UK/2002
10	**IBM** pSeries 690 Turbo 1.3GHz/ 1216	**3164.00** 6323.00	NCAR (National Center for Atmospheric Research) USA/2002
11	**IBM** pSeries 690 Turbo 1.3GHz/ 1184	**3160.00** 6156.00	Naval Oceanographic Office (NAVOCEANO) USA/2002
12	**IBM** SP Power3 375 MHz 16 way/ 6656	**3052.00** 9984.00	NERSC/LBNL USA/2002
13	**IBM** pSeries 690 Turbo 1.3GHz/ 960	**2560.00** 4990.00	ECMWF UK/2002
14	**IBM** pSeries 690 Turbo 1.3GHz/ 960	**2560.00** 4990.00	ECMWF UK/2002
15	**Intel** ASCI Red/ 9632	**2379.00** 3207.00	Sandia National Laboratories USA/1999
16	**IBM** pSeries 690 Turbo 1.3GHz/ 864	**2310.00** 4493.00	Oak Ridge National Laboratory USA/2002
17	**Atipa Technology** P4 Xeon 1.8 GHz - Myrinet/ 1024	**2207.00** 3686.00	Louisiana State University USA/2002

Computing Beyond Teraflop: Grid of Globally Distributed PC Clusters?

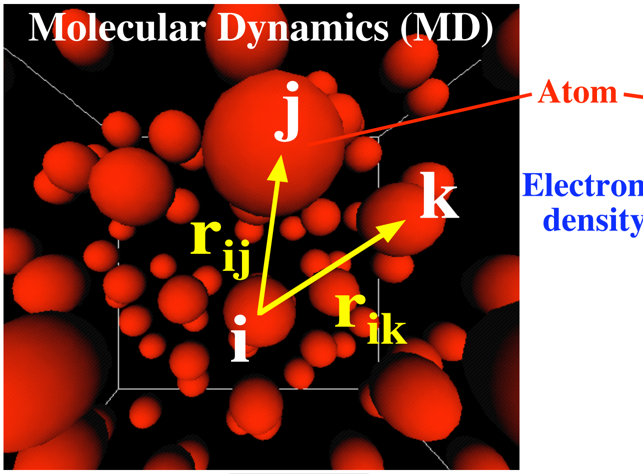


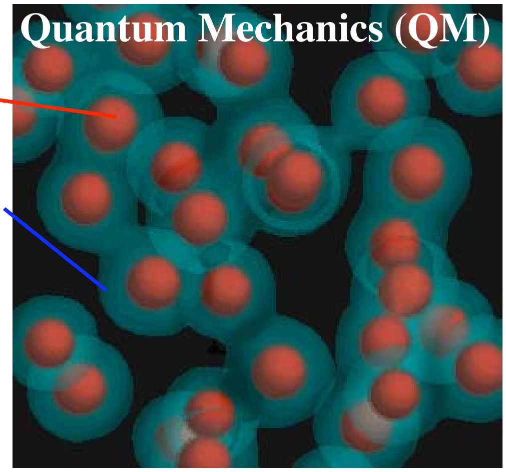
Grid of globally distributed supercomputers

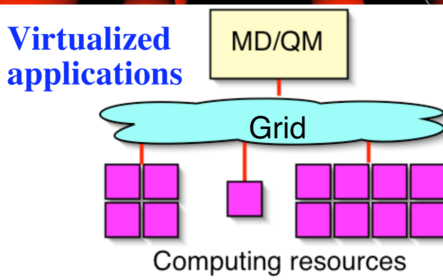
Commodity Xeon-based multi-Teraflop Linux cluster

Virtualization-aware Application Framework

Atomistic materials simulation methods







- **Scalability**
- **Portable performance**
- **Adaptation**

⟶ **Data-locality principles**

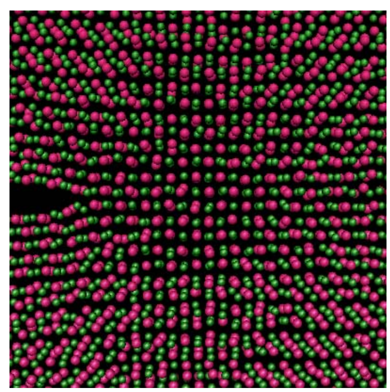
Molecular Dynamics: *N*-Body Problem

- **Newton's equation of motion**

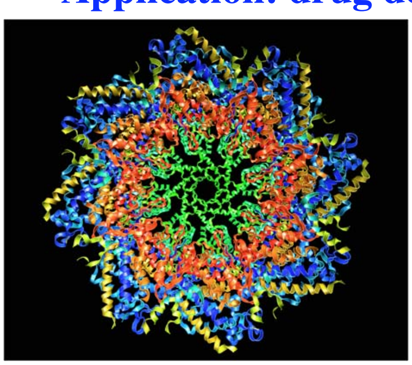
$$m_i \frac{d^2 \mathbf{r}_i}{dt^2} = -\frac{\partial V\left(\mathbf{r}^N\right)}{\partial \mathbf{r}_i} \quad (i = 1, \ldots, N)$$

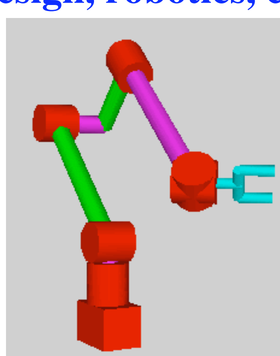
- ***N*-body problem — $O(N^2)$**
  Long-range electrostatic interaction

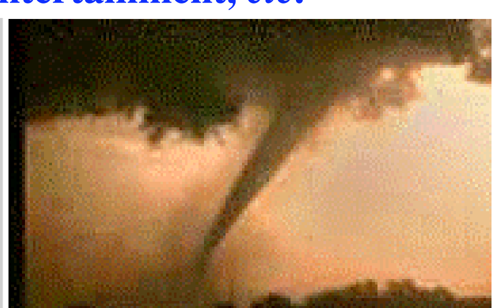
$$V_{es}(\mathbf{x}) = \sum_{i=1}^{N} \frac{q_i}{|\mathbf{x} - \mathbf{x}_i|} \quad \mathbf{x} = \mathbf{x}_j \ (j = 1, \ldots, N)$$



- **Application: drug design, robotics, entertainment, *etc*.**





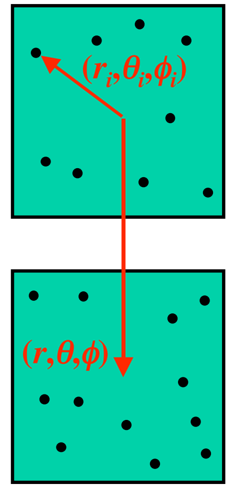


A scene from the movie "Twister"

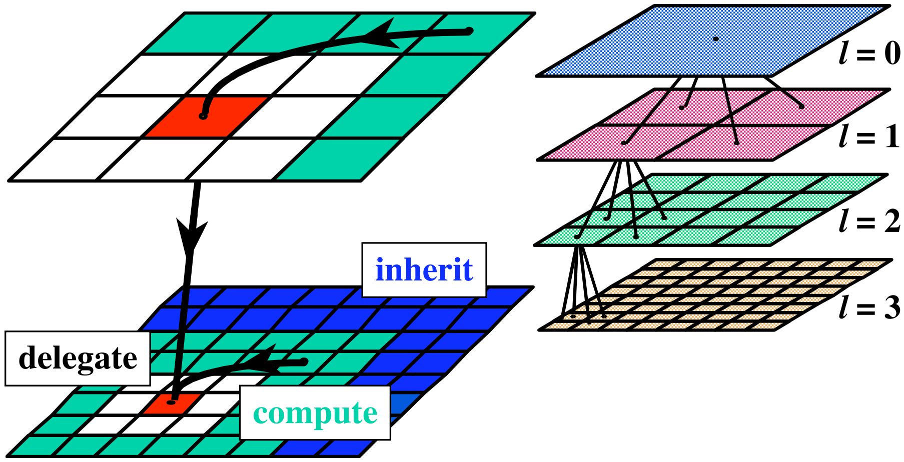
Spatial Locality: Fast Multipole Method

1. **Clustering: Encapsulate far-field information using multipoles**

$$\mathcal{V}(\mathbf{x}) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\left\{\sum_{i=1}^{N} q_i r_i^l Y_l^{*m}(\theta_i,\phi_i)\right\}\frac{Y_l^m(\theta,\phi)}{r^{l+1}}$$

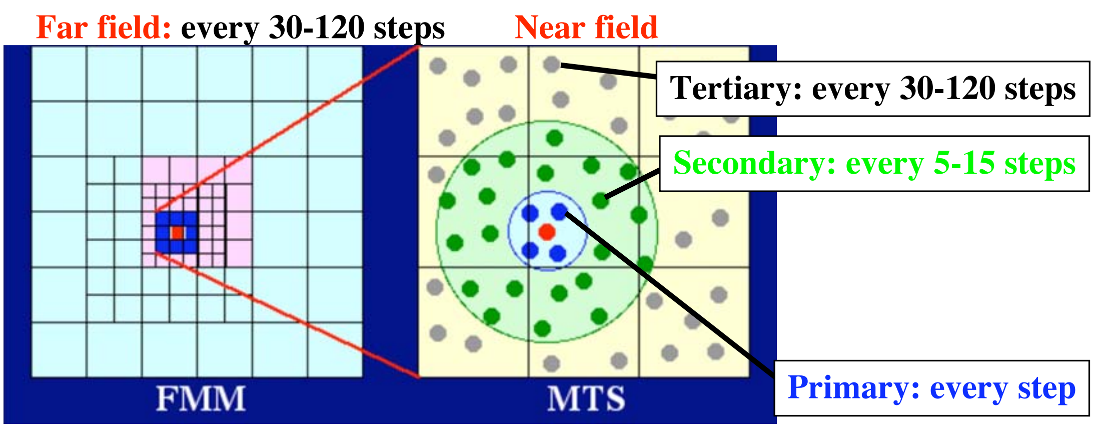


2. **Hierarchical abstraction: Octree data structure**
3. **$O(N)$ algorithm: Constant number of interactive cells per octree node**



Temporal Locality: Multiple Time Stepping

- **Different force-update schedules for different force components**
  → i) Reduced computation
  ii) Enhanced data locality & parallel efficiency

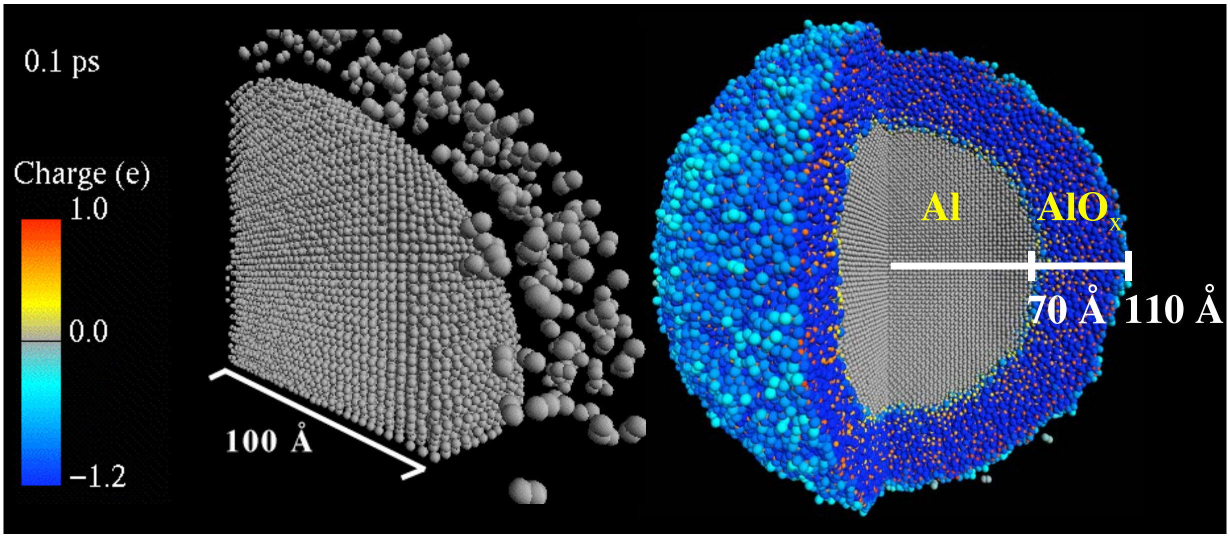


- **Reversible symplectic integrator**
  Simulation-loop invariant: phase-space volume
  → **Long-time stability**

$$\left(\frac{\partial(p^N_{n\Delta t}, r^N_{n\Delta t})}{\partial(p_0^N, r_0^N)}\right)^T \begin{pmatrix}0 & \mathbf{I}\\ -\mathbf{I} & 0\end{pmatrix}\frac{\partial(p^N_{n\Delta t}, r^N_{n\Delta t})}{\partial(p_0^N, r_0^N)} = \begin{pmatrix}0 & \mathbf{I}\\ -\mathbf{I} & 0\end{pmatrix}$$

Oxide Growth in an Al Nanoparticle

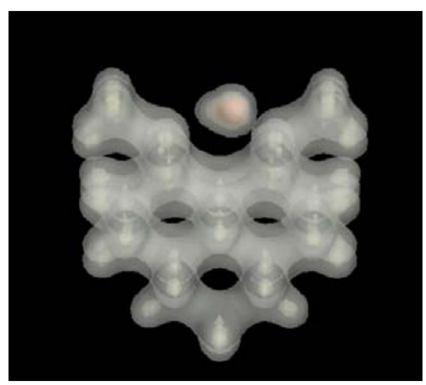
Unique metal/ceramic nanocomposite



Oxide thickness saturates at 40 Å after 0.5 ns —Excellent agreement with experiments

Quantum *N*-Body Problem

Challenge: Exponential complexity



Density functional theory (DFT)

(Kohn, Nobel Chemistry Prize, '98)

$$\psi(r_1, r_2, \ldots, r_{N_{\mathrm{el}}}) \longrightarrow \{\psi_n(\mathbf{r}) \mid n = 1, \ldots, N_{\mathrm{el}}\}$$

$O(C^N)$ $\qquad$ $O(N^3)$

> Pseudopotential **(Troullier & Martins, '91)**

> Generalized gradient approximation **(Perdew, '96)**

Constrained minimization problem:

Minimize:

$$E[\{\psi_n\}] = \sum_{n=1}^{N_{\mathrm{el}}} \int d^3r \psi_n^*(\mathbf{r}) \left(-\frac{\hbar^2}{2m_e} \frac{\partial^2}{\partial \mathbf{r}^2} + V_{\mathrm{ion}}(\mathbf{r}) \right) \psi_n(\mathbf{r}) + \frac{e^2}{2} \iint d^3r d^3r' \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} + E_{\mathrm{XC}}[\rho(\mathbf{r})]$$

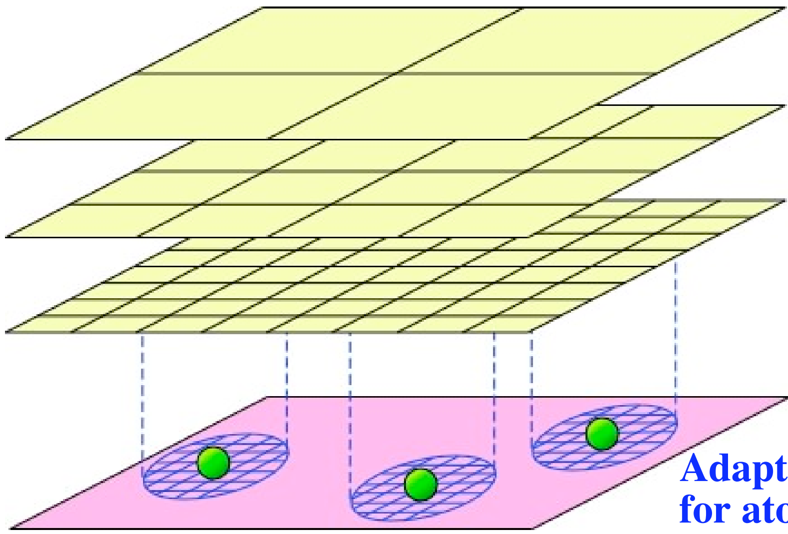
with orthonormal constraints: $\int d^3r \psi_m^*(\mathbf{r}) \psi_n(\mathbf{r}) = \delta_{mn}$

Charge density: $\rho(\mathbf{r}) = \sum_{n=1}^{N_{\mathrm{el}}} |\psi_n(\mathbf{r})|^2$

Real-Space DFT on Hierarchical Grids

Efficient parallelization of DFT: real-space approaches

- **High-order finite difference** (Chelikowsky, Troullier, Saad, '94)
- **Multigrid acceleration** (Bernholc *et al.*, '96; Beck, '00)
- **Double-grid method** (Ono, Hirose, '99)
- **Spatial decomposition/divide-&-conquer**



Multigrid acceleration of preconditioned conjugate gradient

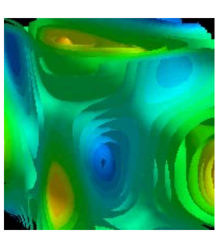
Adaptive high-resolution grid for atomic pseudopotentials

Quantum-Nearsightedness Locality Principle

***O*(*N*) DFT algorithm** (Mauri & Galli, '94)

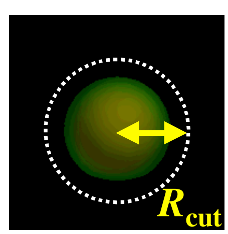
- **Asymptotic decay of density matrix:**
- **Localized functions:**

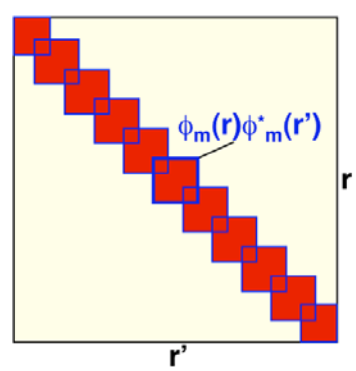
$$\rho(\mathbf{r},\mathbf{r}') \equiv \sum_{n=1}^{N_{\rm el}} \psi_n(\mathbf{r})\psi_n^*(\mathbf{r}') \propto \exp(-C|\mathbf{r}-\mathbf{r}'|)$$



$\psi_n(\mathbf{r}) \rightarrow \phi_m(\mathbf{r})$

$$\phi_m(\mathbf{r}) = \sum_n \psi_n(\mathbf{r})U_{nm}$$





$\rho(\mathbf{r},\mathbf{r}')$

- **Unconstrained minimization:**

$$\tilde{E}[\{\phi_n\}] = \sum_{m=1}^{N_{\rm wf}} \sum_{n=1}^{N_{\rm wf}} \int d^3r\phi_m^*(\mathbf{r})(H-\eta I)\phi_n(\mathbf{r})\left(2\delta_{nm} - \int d^3r\phi_n^*(\mathbf{r})\phi_m(\mathbf{r})\right) + \eta N_{\rm el}$$

Data Locality in Parallelization

Challenge: Load balancing for irregular data structures

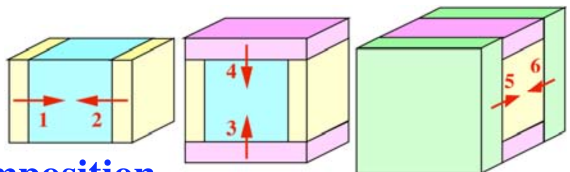
Irregular data-structures/ processor-speed

Map →

Parallel computer

Optimization problem:

- Minimize the load-imbalance cost
- Minimize the communication cost
- Topology-preserving spatial decomposition
  → structured 6-step message passing minimizes latency



$$E = t_{\mathrm{comp}}\left(\max_p \left|\{i \mid \mathbf{r}_i \in p\}\right|\right) + t_{\mathrm{comm}}\left(\max_p \left|\{i \mid \|\mathbf{r}_i - \partial p\| < r_c\}\right|\right) + t_{\mathrm{latency}}\left(\max_p \left[N_{\mathrm{message}}(p)\right]\right)$$

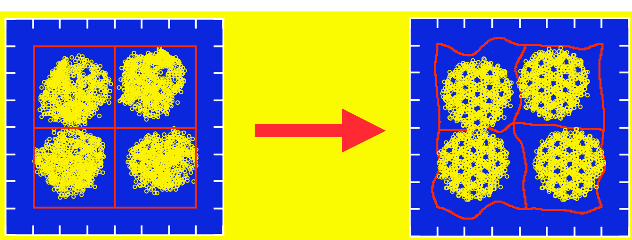
Computational-Space Decomposition

Topology-preserving "computational-space" decomposition in curved space

Curvilinear coordinate transformation
$\xi = \mathbf{x} + \mathbf{u}(\mathbf{x})$

Particle-processor mapping: regular 3D mesh topology

$$\begin{cases} p(\xi_i) = p_x(\xi_{ix})P_yP_z + p_y(\xi_{iy})P_z + p_z(\xi_{iz}) \\ p_\alpha(\xi_{i\alpha}) = \lfloor \xi_{i\alpha}P_\alpha / L_\alpha \rfloor \quad (\alpha = x,y,z) \end{cases}$$

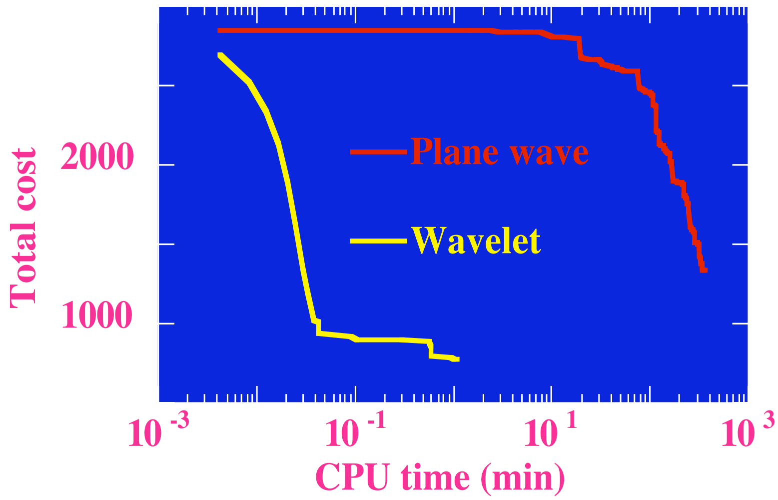


Regular mesh topology in computational space, ξ — Curved partition in physical space, x

Wavelet-based Adaptive Load Balancing

- Simulated annealing to minimize the load-imbalance & communication costs, $E[\xi(x)]$
- Wavelet representation speeds up the optimization

$$\xi(x) = x + \sum_{l,m} d_{lm} \psi_{lm}(x)$$

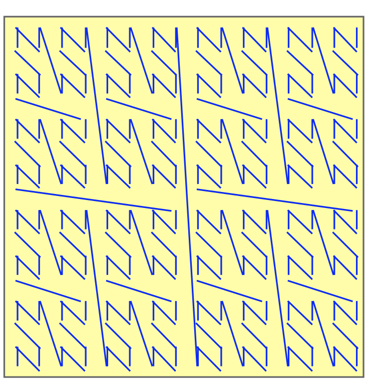


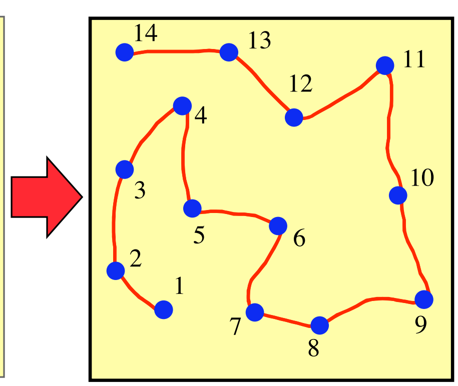
Locality in Data Compression

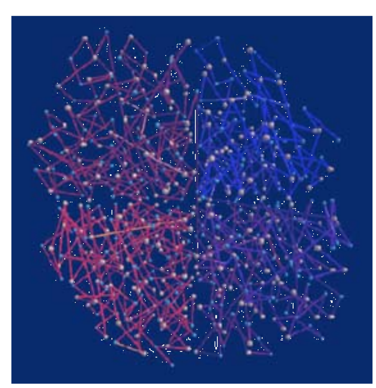
Massive data transfer via wide area network:
75 GB/step of data for 1.5 billion-atom MD!
→ Compressed software pipeline

Scalable encoding:
- **Store relative positions on spacefilling curve: $O(N\log N) \rightarrow O(N)$**

Result:
- **Data size, 50 Bytes/atom → 6 Bytes/atom**





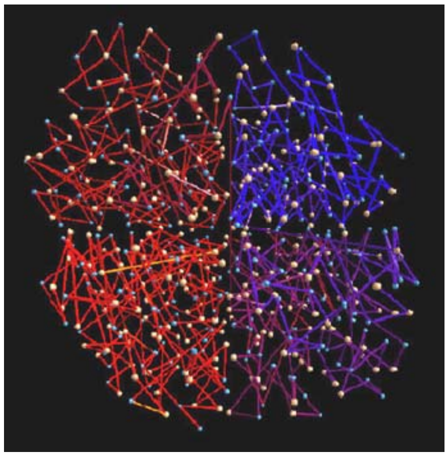


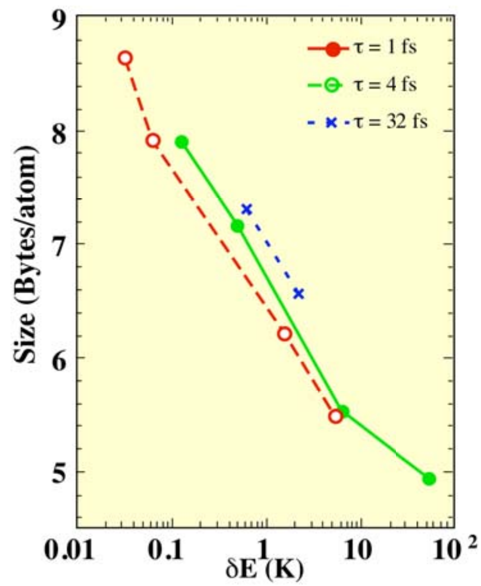
Spacefilling-curve Data Compression

Algorithm:

1. Sort particles along the spacefilling curve
2. Store relative positions: $O(N\log N) \rightarrow O(N)$

- Adaptive variable-length encoding to handle outliers
- User-controlled error bound



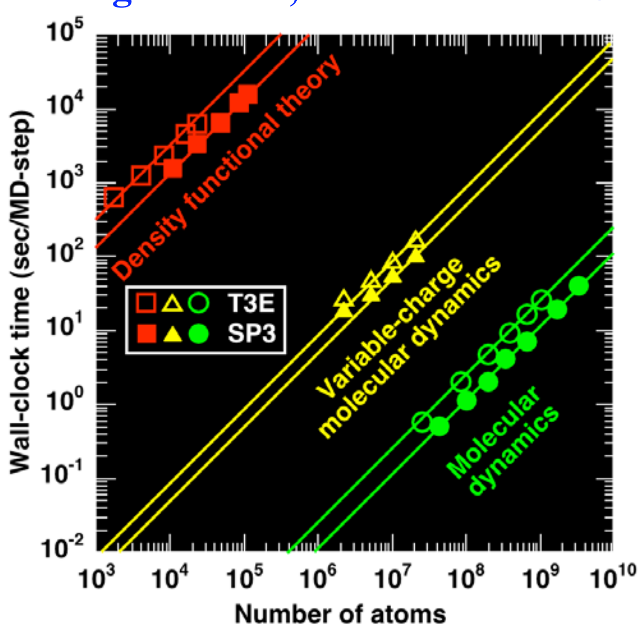


Result:

- An order-of-magnitude reduction of I/O size: 50 → 6 Bytes/atom

Scalable Scientific Algorithm Suite

Design-space diagram on 1,024 IBM SP3 & Cray T3E processors



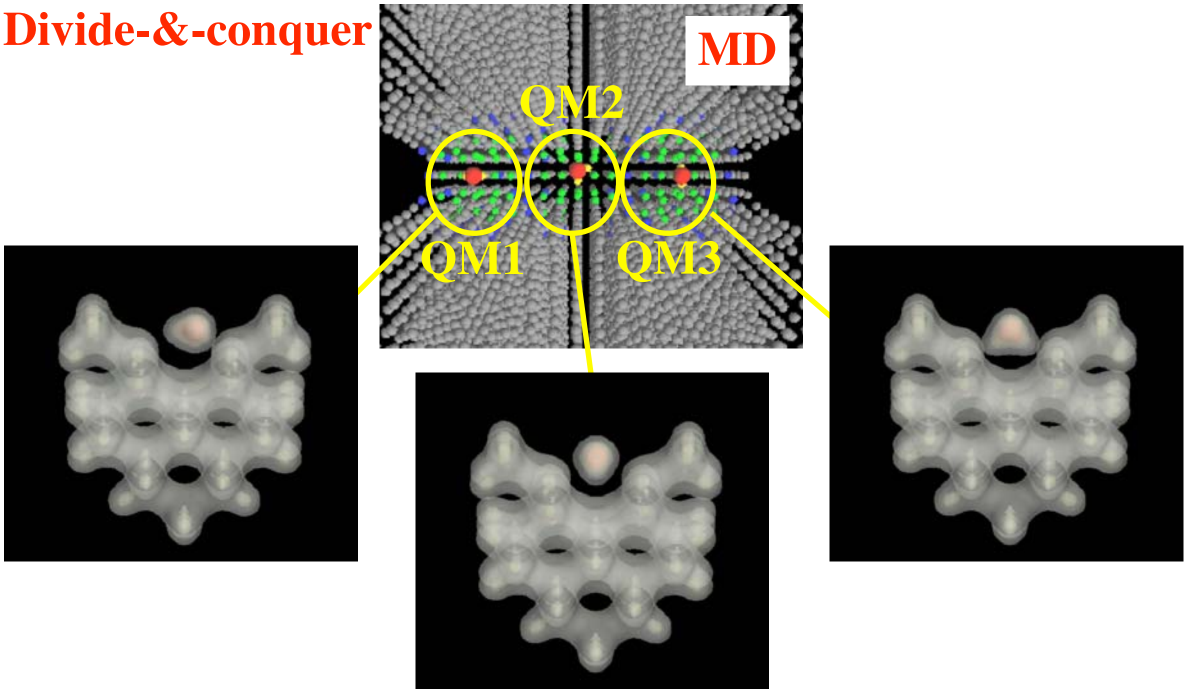
On 1,024 IBM SP3 processors:
- 6.44-billion-atom MD of SiO_2
- 444,000-electron DFT of GaAs

Supercomputing 2001 Best Technical Paper Award

Grid Enabling: Multiple QM Clustering

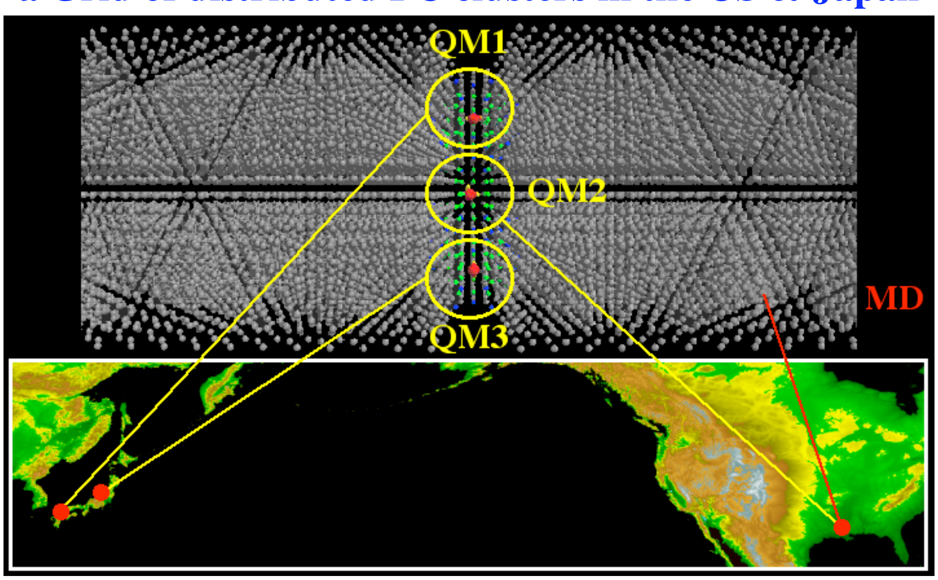
$$E = E_{\mathrm{MD}}^{\mathrm{system}} + \sum_{\mathrm{cluster}} [E_{\mathrm{QM}}^{\mathrm{cluster}}(\{\mathbf{r}_{\mathrm{QM}}\},\{\mathbf{r}_{\mathrm{HS}}\}) - E_{\mathrm{MD}}^{\mathrm{cluster}}(\{\mathbf{r}_{\mathrm{QM}}\},\{\mathbf{r}_{\mathrm{HS}}\})]$$

Divide-&-conquer



Global Collaborative Simulation

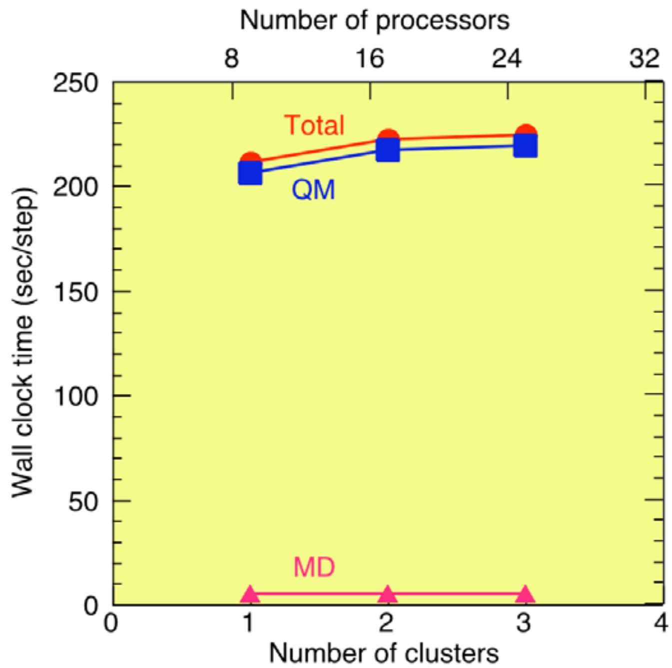
Hybrid MD/QM simulation on
a Grid of distributed PC clusters in the US & Japan



Japan: Yamaguchi — 65 P4 2.0GHz
Hiroshima, Okayama, Niigata — 3×24 P4 1.8GHz

US: Louisiana — 17 Athlon XP 1900+

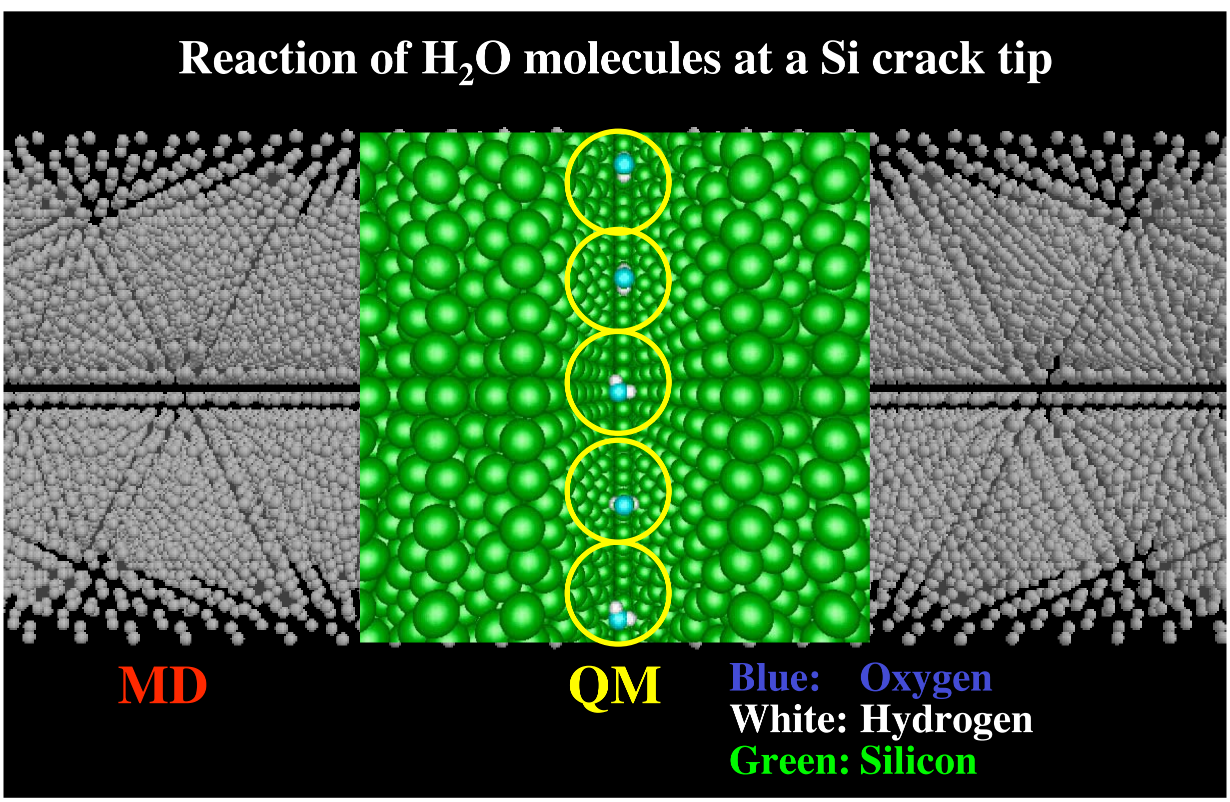
Preliminary Benchmark Results



MD
91,256 atoms
QM (DFT)
76*n* atoms on *n* clusters

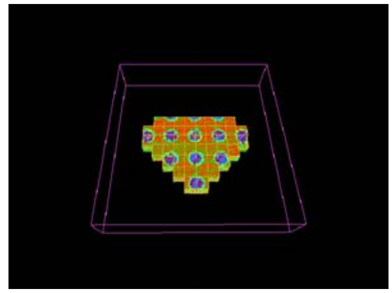
- **Scaled speedup, *P* = 1 (for MD) + 8*n* (for QM)**
- **Efficiency = 94.0% on 25 processors over 3 PC clusters in the US & Japan**

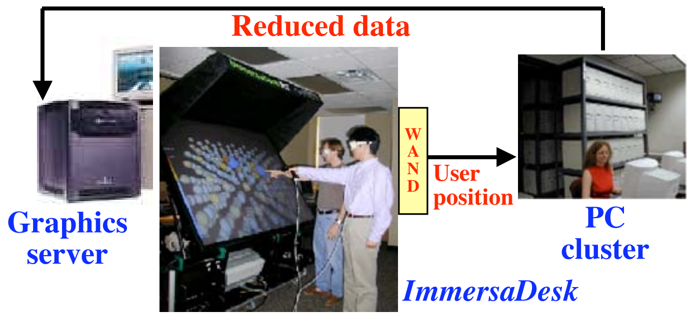
Environmental Effect on Fracture

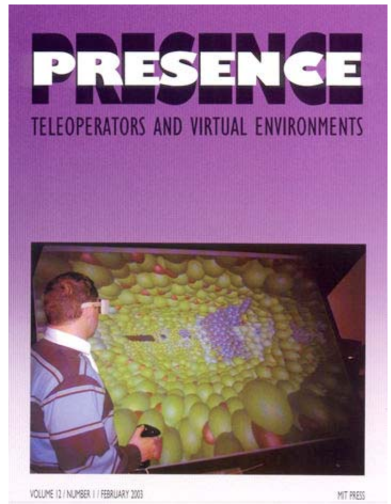


Data Locality in Visualization

- Octree-based fast view-frustum culling
- Probabilistic occlusion culling
- Parallel/distributed processing

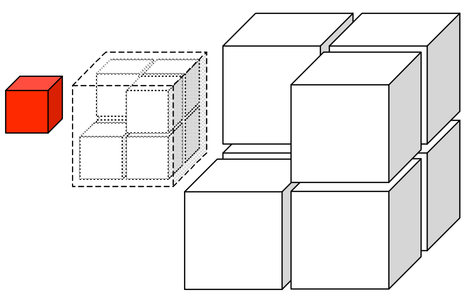


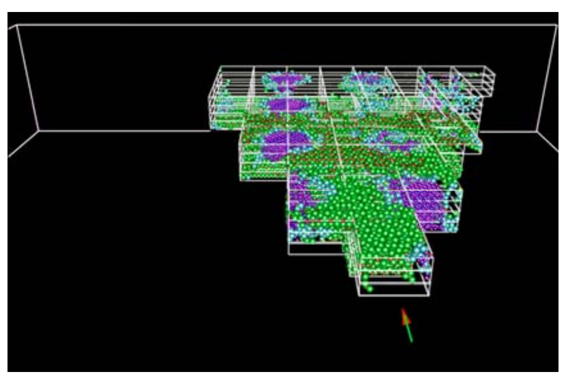


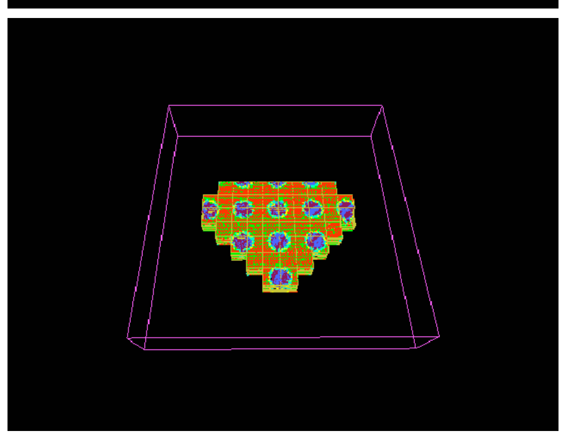


- Interactive visualization of a billion-atom dataset in immersive environment

Octree-based View-Frustum Culling



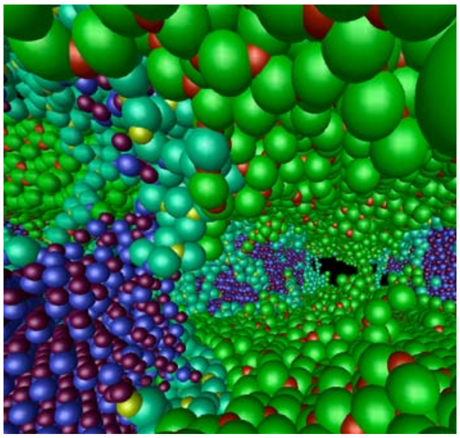




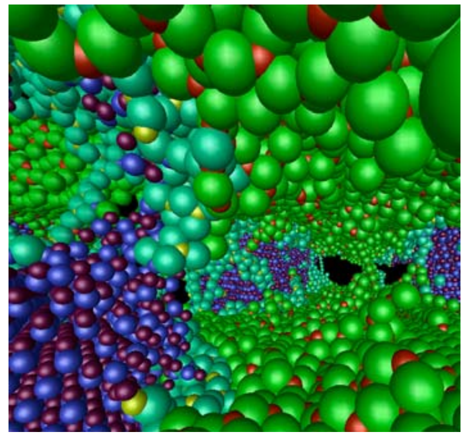
- Use the octree data structure to efficiently select only visible atoms
- Complexity
  Insertion into octree: $O(N)$
  Data extraction: $O(\log N)$

Probabilstic Occlusion Culling

- Remove atoms that are occluded by other atoms closer to the viewer
- Draw fewer atoms per region as the distance of a region from the viewer increases

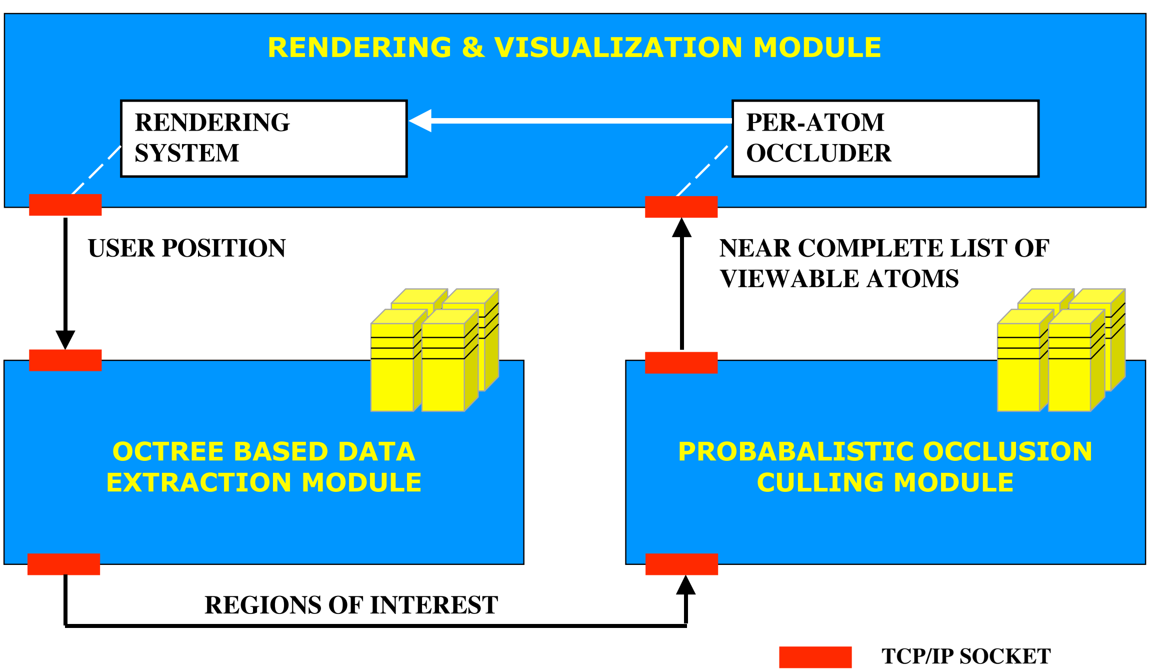


Without occlusion



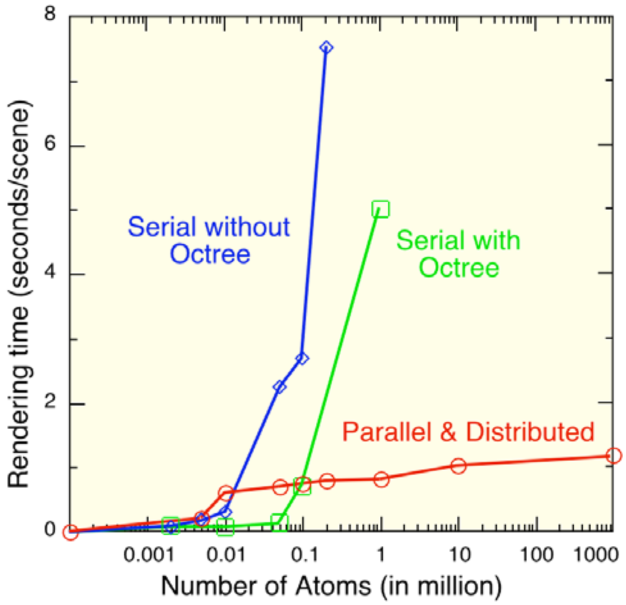
With occlusion
68% fewer atoms &
3 times higher frame rate

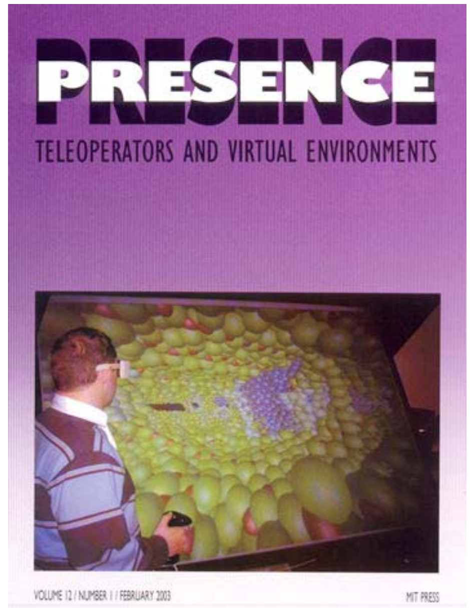
Distributed Architecture



Parallel & Distributed Atomsviewer

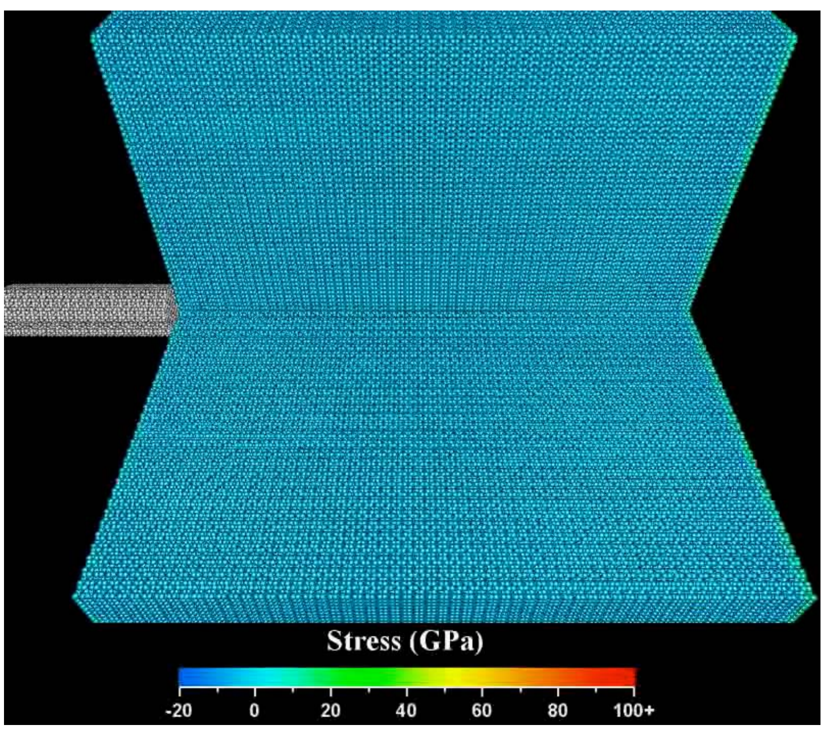
Real-time walkthrough for a billion atoms on an SGI Onyx2 (2 × MIPS R10K, 4GB RAM) connected to a PC cluster (4 × 800MHz P3)





IEEE Virtual Reality 2002 Best Paper

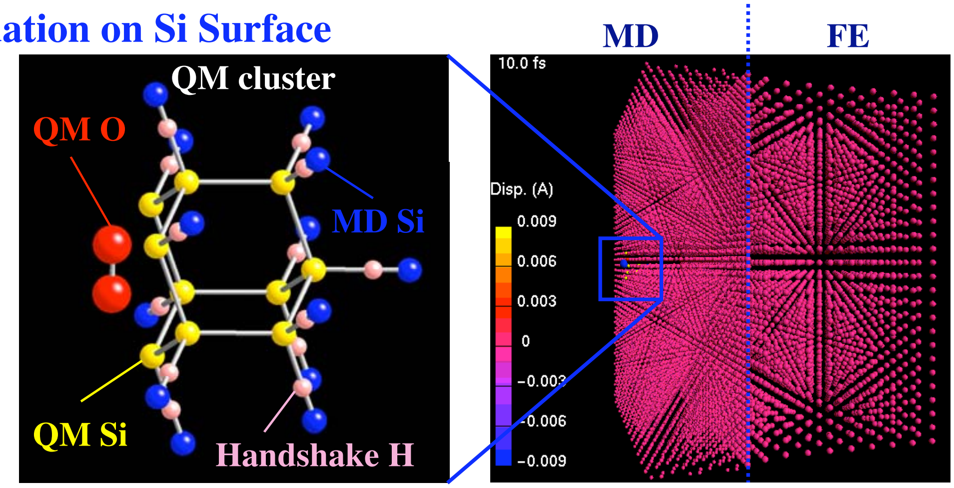
209 Million Atom MD of Hypervelocity Impact



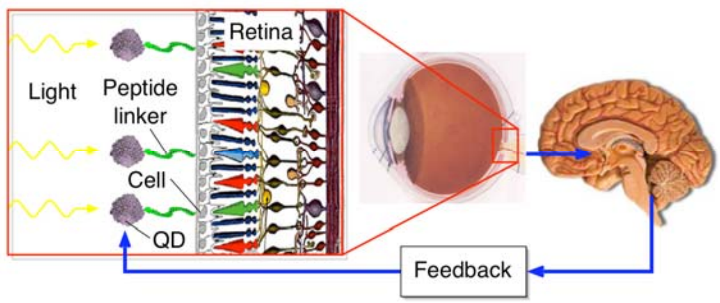
AlN plate with impact velocity 15 km/s

Application of Multiscale Simulations

- Oxidation on Si Surface



- Interfacing quantum-dot devices & biological cells
  (*e.g.* neural implant to restore vision)

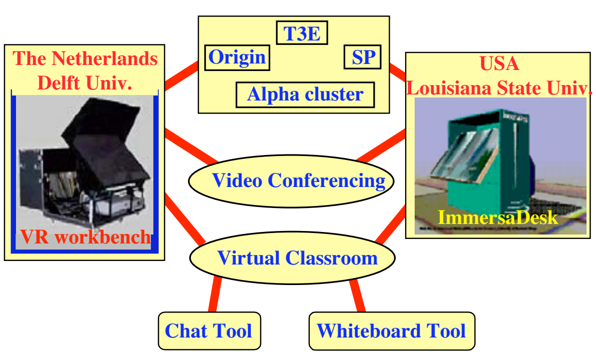


Computational Science Education

Dual-Degree Program

Ph.D. in physical/biological sciences & MS from computer science

Intercontinental Courses



Computational Science Workshop for Underrepresented Groups

Conclusion

Multiscale simulation approach:

- **Can be virtualized on a Grid of distributed PC clusters through data-locality principles**
- **Will allow global collaboration of scientists to increase the scope & size of simulation study**