



Parallel Computing and Many Body Problems

George Biros Bastiaan Braams

Class info

- Office hours
 - G. Biros : Monday 4-6pm
 - B. Braams: Tuesday 4-6pm

- Class home page
 - <http://cs.nyu.edu/courses/spring03/G22.2945-001/index.htm>

Class info - Requirements

- Homeworks
 - Algorithm design
 - Shared memory programming
 - MPI programming
- Semester project

Class Info - Topics

- Parallel computing
 - Algorithmic primitives
 - Shared memory
 - Distributed memory
- Scientific computing
 - N-Body algorithms
 - Multigrid, FFT
 - Linear Algebra
- Statistical physics
 - Monte Carlo simulations
 - Ising model

Class info – computing

- Six 4 CPU Sun workstations
Bionum{1,6}.cims.nyu.edu
- One 8 CPU SGI Origin,
spectrum.cims.nyu.edu
- Two 4 CPU SGI Origin,
{septum,stratum}.cims.nyu.edu

Introduction

Why parallel computing?

1. Emergence of Computational Science

- Science
 - Analysis
 - Restricted to model problems, simple geometries
 - Experiments
 - Expensive (crash worthiness, aerodynamics)
 - Impossible (Astrophysics, Earthquakes, Global climate)
 - Dangerous (Medical devices, Nuclear devices)
 - Difficult to reproduce
- New Venue: Computational Science
 - Direct modeling of physical phenomena for
 - Scientific discovery
 - Optimal design
 - Engineering and Industry

Why parallel computing?

2. Sequential computing is slow

- To get faster has to get smaller, but
- Physics limitations
- Parallel is faster by definition

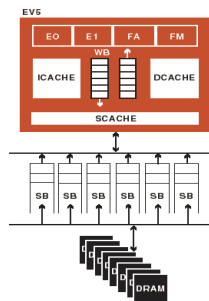
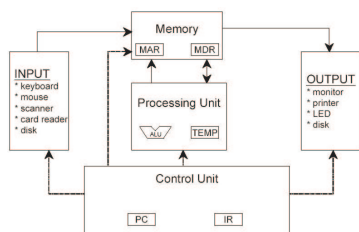
- But tuning software to run fast in single CPU
- is very important

Earth simulator – Fastest (silicon-based) machine as of 2002, 40Tflops ~20,000 P4s



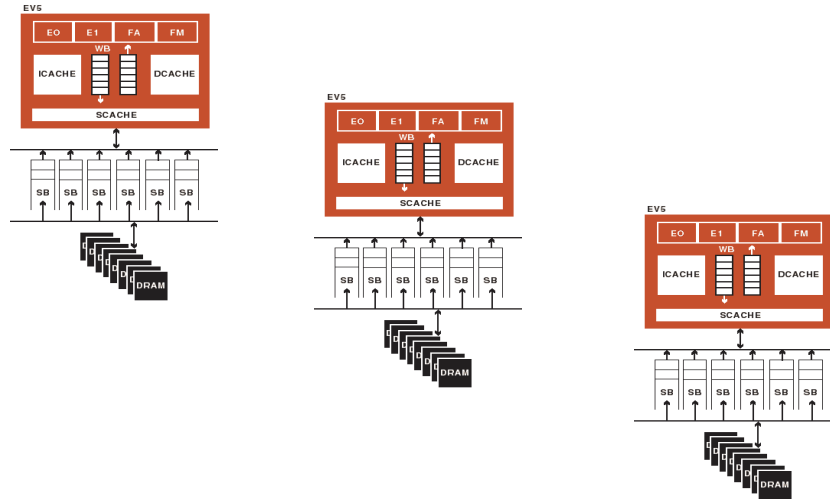
Von Neumann computing model

Memory, CPU, I/O



In practice several memory hierarchies
Rule: Large memory is slow
Small memory is fast

Parallelism



Basic Definitions

- Speedup
 - best sequential / time on p processors
- Efficiency
 - Speedup/ p (< 1)
- Latency
 - time to initiate communication channel
- Bandwidth
 - capacity of communication channel

Efficiency

- Algorithmic scalability (sequential complexity)
 - How the algorithm scales with the increasing problem size and fixed number of processors
- Architecture scalability (fixed size scalability)
 - How the algorithm scales with fixed problem size and fixed number of processors
- Overall scalability (iso-granular scalability)
 - Fixed grain size (work per processor). Both work and p increase. The most important in applications

Amdahl's law

- Sequential bottleneck will ruin scalability
- s is the sequential part percentage on the overall work.

$$E = \frac{1}{s + (1 - s)/p} \leq \frac{1}{s}$$

- Fix: sequential part should be independent of problem size, and s will decrease as problem becomes bigger.

Basic definitions - continued

- Coordination
 - Synchronous vs. Asynchronous
- Scalability
 - Number of processors
- Granularity
 - Single processor work
- Interconnection network (p^2 is too expensive)
 - Ring, Bus, Mesh, Torus, Star, Hypercube, Butterfly, Fat trees
- Memory
 - Registers, Cache (L1, L2, L3), RAM, Discs

Basic models

- Machine models
 - Single Instruction Single Data (SISD)
 - Data parallel (Vector) (SIMD)
 - Shared memory (SMP)
 - Distributed memory (MIMD, SPMD)
- Programming models
 - Compilers (HPF, HPC++)
 - Threads, OpenMP
 - Message Passing MPI, PVM
- Best platforms combine everything (SMP Clusters)
- "Efficient" software should combine OpenMP + MPI

Basic steps in writing programs

- Partition work
- Determine communication
- Agglomeration to number of available processors
- Map to processors

- Goals
 - Minimize communication
 - Maximize concurrency of communication
 - Minimize synchronizations
 - Overlap computation with communication
 - Load balance
 - Avoid Amdahl law (sequential part that scales with input size)

Basic work partitioning techniques

- Divide and conquer
 - Important applications in N-Body algorithms
- Pipelining
 - Overlapping similar computation phases
- Domain decomposition
 - Partition of work is based on input data
- Functional decomposition
 - Partition is based on computation
- Embarassingly parallel
 - Independent tasks are readily identified

Models for algorithm evaluation

- Work/Depth models
 - Vector, Language, Graphs
- PRAM (shared memory)
 - Access to memory takes unit time
 - Variants to support exclusive reads and writes
- BSP - Bulk, Synchronous, Parallel (distributed)
 - Local/Remote memory
 - Uniform times to access remote memory
- LogP
 - Latency, Overhead, Gap (communication bandwidth), Processors

Practical goals

- Numerical Algorithms must be
 - highly concurrent and straightforward to load balance
 - latency tolerant
 - cache friendly (temporal and spatial locality of reference)
 - highly scalable (in the sense of algorithm convergence)
- Goal for algorithmic scalability: fill up memory of arbitrarily large machines **while preserving constant running times** with respect to proportionally smaller problem on one processor

Importance of optimal algorithms

- M1 runs in $O(N^2)$ in 1 CPU
- M2 runs in $O(N)$ in 1 CPU

- On 1000 CPUs M1 solves problem of size 30 x N
- On 1000 CPUs M2 solves problem of size 1000 x N

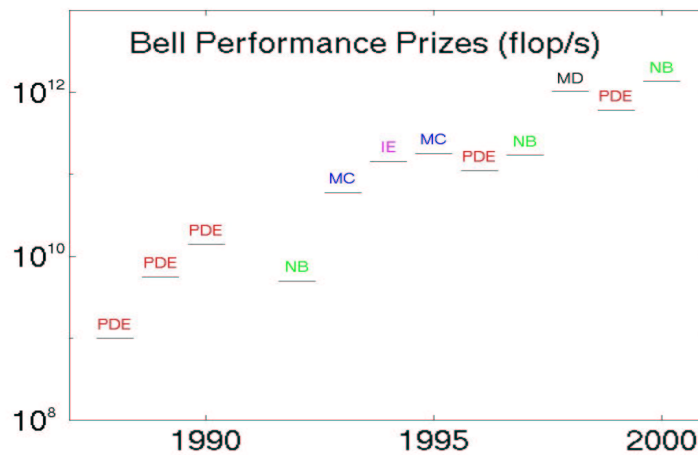
Software engineering

- Programming is more difficult
 - Deadlocks
 - System level requirements
 - Check-pointing
- Unreliable OS
- Unreliable I/O
- Unreliable software
- Unreliable hardware
- Debugging is painful

Things have changed

- Robust platforms for “small” p
 - 1-256 CPU. Shared memory, SMPs and Beowulf clusters
 - MPI libraries *de facto* standard
 - Portability is possible
 - Debugging, performance monitoring, and software libraries readily available

Gordon Bell Prize winners



Bell prizes history

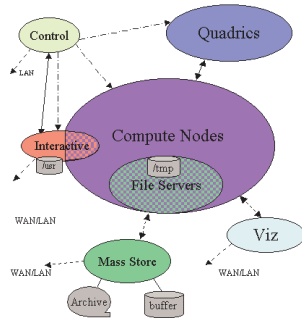
<i>Year</i>	<i>Type</i>	<i>Application</i>	<i>Gflop/s</i>	<i>System</i>	<i>No. Procs</i>
1988	PDE	Structures	1.0	Cray Y-MP	8
1989	PDE	Seismic	5.6	CM-2	2,048
1990	PDE	Seismic	14	CM-2	2,048
1992	NB	Gravitation	5.4	Delta	512
1993	MC	Boltzmann	60	CM-5	1,024
1994	IE	Structures	143	Paragon	1,904
1995	MC	QCD	179	NWT	128
1996	PDE	CFD	111	NWT	160
1997	NB	Gravitation	170	ASCI Red	4,096
1998	MD	Magnetism	1,020	T3E-1200	1,536
1999	PDE	CFD	627	ASCI BluePac	5,832
2000	NB	Gravitation	1,349	GRAPE-6	96

2002 Gordon Bell awards www.supercomp.org

- Astrophysics* (26.5 Tflops)
 - GRAPE 6, Japan
 - Structural mechanics (1.16 Tflops 3K CPUs)
 - Salinas, Sandia Labs
 - Atmospheric simulations (26.5 Tflops 5K CPUs)
 - Earth simulator, Japan
 - Turbulence simulations (16.4 Tflops)
 - Earth simulator, Japan
 - Nucluar Fusion
 - Earth simulator, Japan (14.9 Tflops)
 - Biomolecular simulation*
 - NAMD, UIUC
- *N-body algorithm

Fastest academic computer in US (2002)

Terascale Computing System

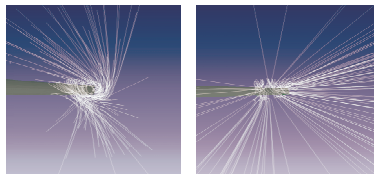


Summary

- 750 Compute Nodes
- 3000 EV68 processors
- 6 Tf (peak, est >4Tf on LSMS)
- 3.0 TB memory
- 40 TB local disk (sys + tmp)
- Multi-rail fat-tree network
- Redundant monitor/ctrl
- WAN/LAN accessible
- Parallel visualization
- File servers: 30TB, ~32 GB/s
- Mass store, ~1 TB/hr

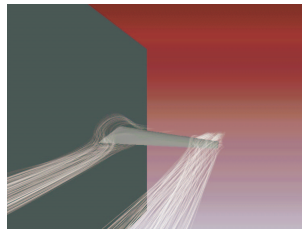


Control of flow around a Boeing 707 wing



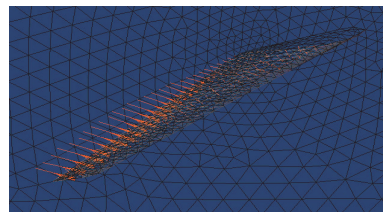
uncontrolled

controlled



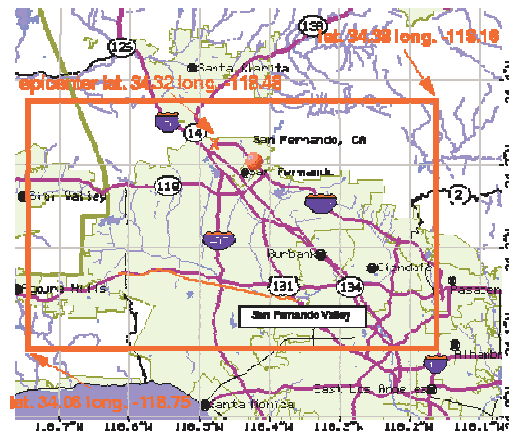
Optimal control of laminar viscous flow

- optimization variables are surface suction/injection
- objective is minimum drag
- 700,000 states; 4,000 controls
- 128 Cray T3E processors
- ~5 hrs for optimal solution (~1 hr for analysis)



Suction/Injection control

Simulation of a 1994 Earthquake at Northridge

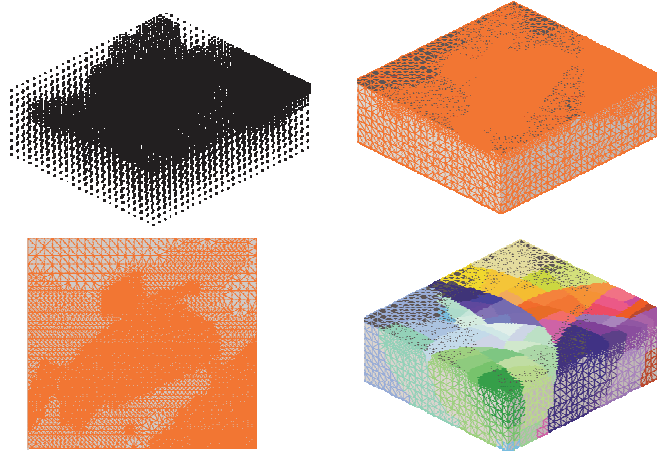


Efficiencies and required memory

PEs	model	grids pts	grid pts/PE	total Gflops/s	Mflops/s/PE	efficiency
1	LA10	134,500	134,500	0.506	506	1.00
16	LA5	618,672	38,667	7.85	490	0.970
64	LA2	7,934,272	123,973	30.7	479	0.947
512	LA1	47,556,096	92,883	231	451	0.891
1024	LA1H	101,940,152	99,551	450	439	0.868
3000	LA0.5	300,000,000	100,000			

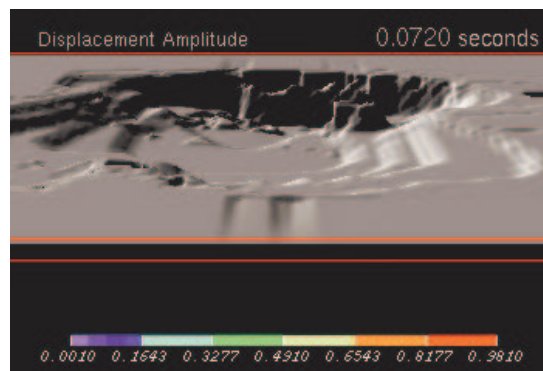
- Largest current (complete) simulation:
 - 1Hz source takes 6h on 512PEs @ 231 Gflops/s
- Target simulation:
 - 2Hz, estimated 13.5h on 3000 PEs @ 1.2 Tflops/s

Unstructured grids and partitioning

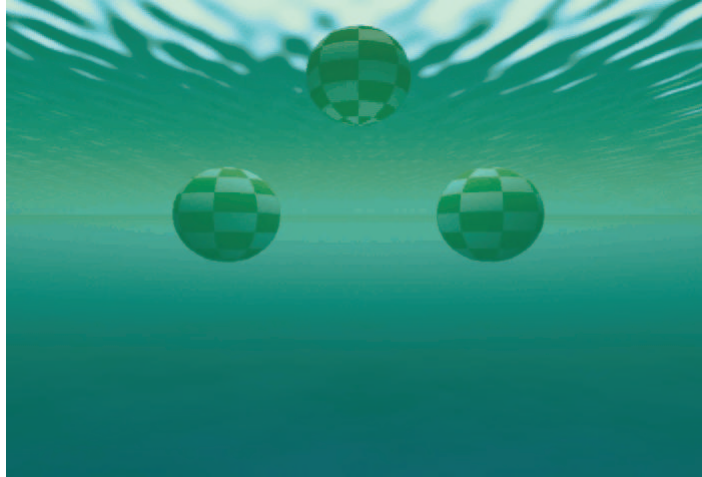


Simulation, 100^6 elements Cray T3E at PSC

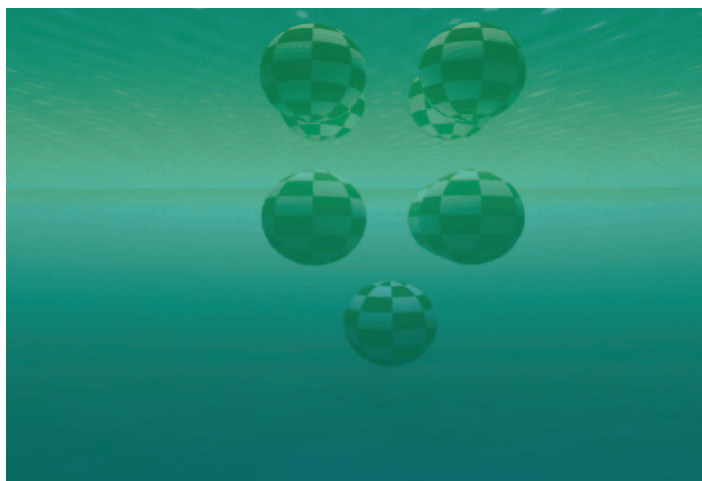
www.cs.cmu.edu/~oghattas



N-Body Algorithms- Integral equations



N-Body algorithms – Integral equations



Top 500

www.top500.org

Rank	Manufacturer Computer / Procs	R _{max} R _{min}	Installation Site Country / Year
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