GPU Computing

John Stone
Theoretical and Computational Biophysics Group
University of Illinois at Urbana-Champaign
http://www.ks.uiuc.edu/Research/gpu/

Cape Linux Users Group, October 28, 2008

Evolution of Graphics Hardware Towards Programmability

- As graphics accelerators became more powerful, an increasing fraction of the graphics processing pipeline was implemented in hardware
- For performance reasons, this hardware was highly optimized and task-specific
- Over time, with ongoing increases in circuit density and the need for flexibility in lighting and texturing, graphics pipelines gradually incorporated programmability in specific pipeline stages
- Modern graphics accelerators are now processors in their own right (thus the new term “GPU”), and are composed primarily of large arrays of programmable processing units

Programmable Graphics Hardware

Groundbreaking research systems:

AT&T Pixel Machine (1989):

82 x DSP32 processors

UNC PixelFlow (1992-98):

64 x (PA-8000 +

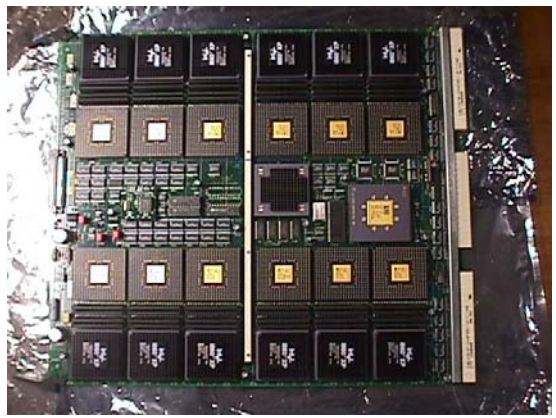
8,192 bit-serial SIMD)

SGI RealityEngine (1990s):

Up to 12 i860-XP processors perform vertex operations (*u*code), fixed-func. fragment hardware

All mainstream GPUs now incorporate programmable processors

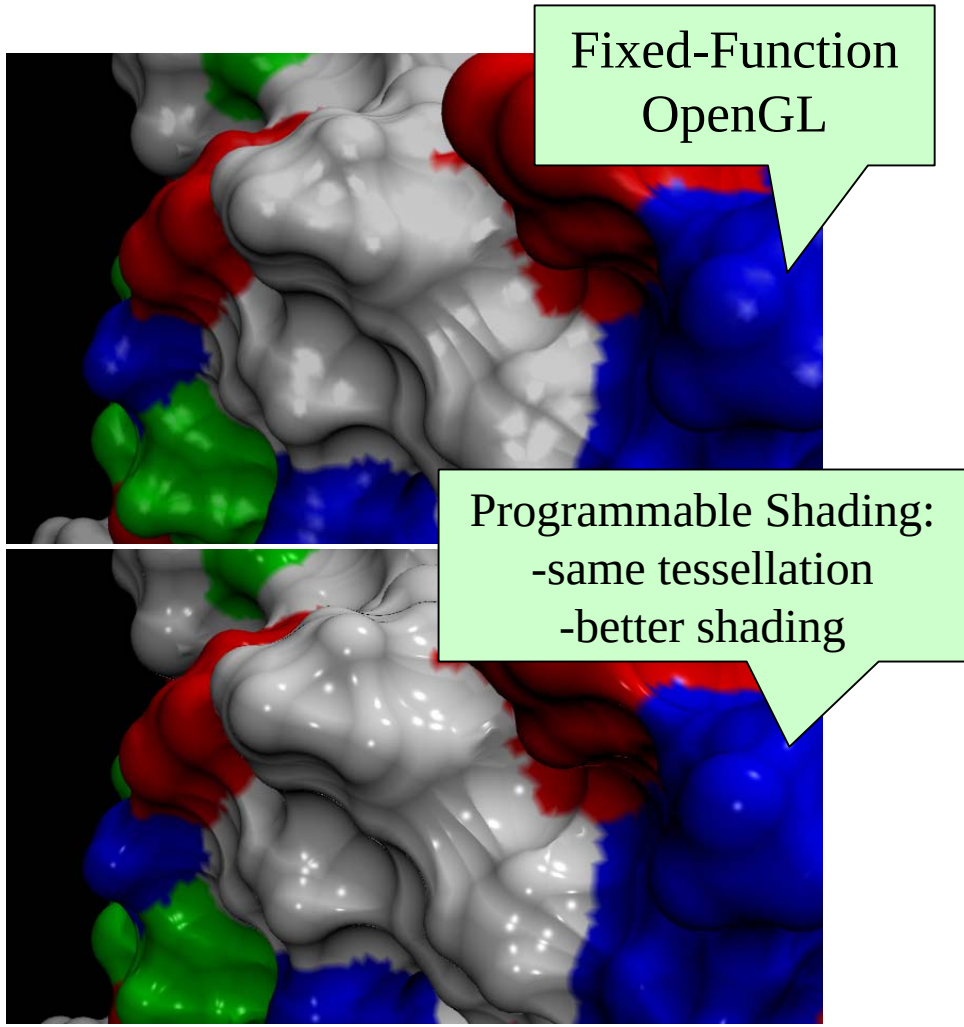
UNC PixelFlow Rack



Reality Engine Vertex Processors

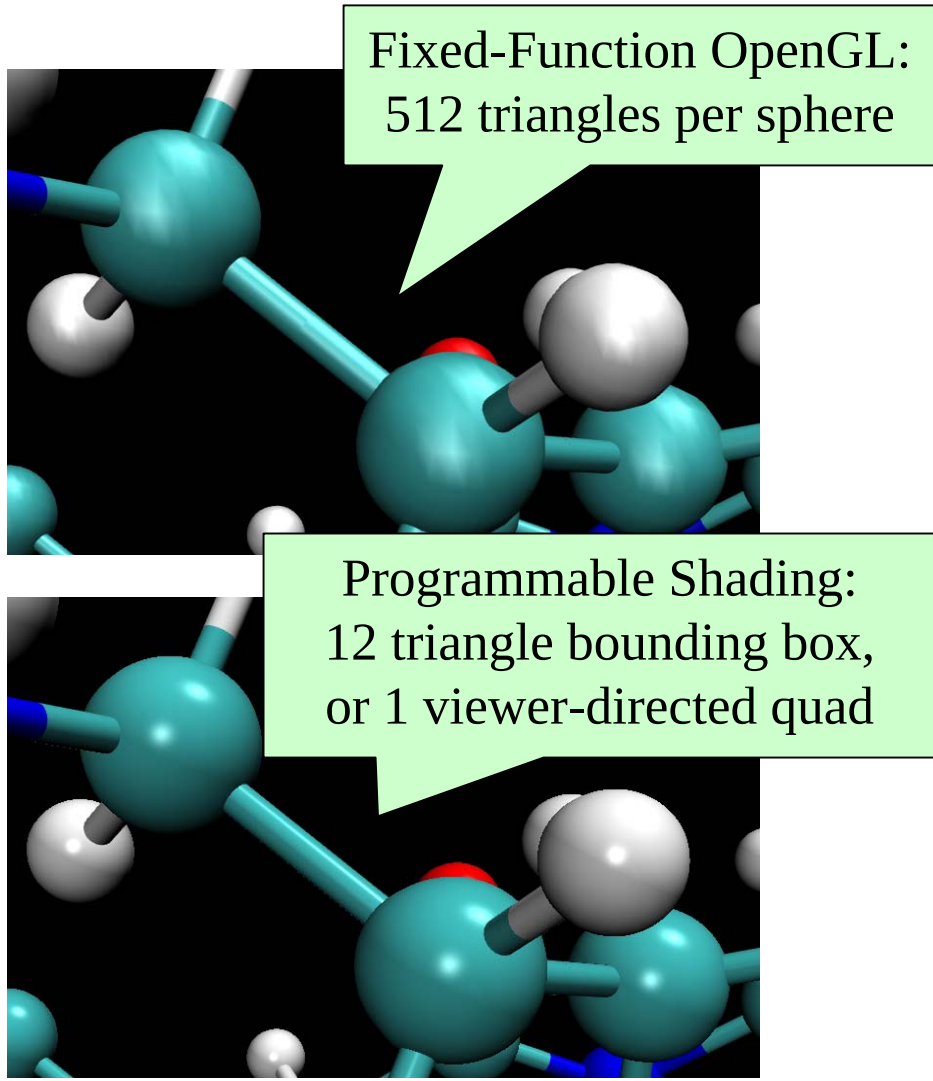
Benefits of Programmable Shading

- Potential for superior image quality with better shading algorithms
- Direct rendering of:
  - Quadric surfaces
  - Volumetric data
- Offload work from host CPU to GPU

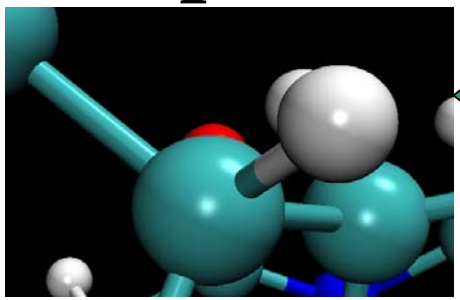


Ray Traced Sphere Rendering with Programmable Shading

- Fixed-function OpenGL requires curved surfaces to be tessellated with triangles, lines, or points
- Fine tessellation required for good results with Gouraud shading; performance suffers
- Static tessellations look bad when viewer zooms in
- Programmable shading solution:
  - Ray trace spheres in fragment shader
  - GPU does all the work
  - Spheres look good at all zoom levels
  - Rendering time is proportional to pixel area covered by sphere
  - Overdraw is a bigger penalty than for triangulated spheres



Sphere Fragment Shader



- Written in OpenGL Shading Language
- High-level C-like language with vector types and operations
- Compiled dynamically by the graphics driver at *runtime*
- Compiled machine code executes on GPU

```
//
// VMD Sphere Fragment Shader (not for normal geometry)
//
void main(void) {
  vec3 raydir = normalize(V);
  vec3 spheredir = spherepos - rayorigin;

  // Perform ray-sphere intersection tests based on the code in Tachyon
  float b = dot(raydir, spheredir);
  float temp = dot(spheredir, spheredir);
  float disc = b*b + sphereradsq - temp;

  // only calculate the nearest intersection, for speed
  if (disc <= 0.0)
    discard; // ray missed sphere entirely, discard fragment

  // calculate closest intersection
  float tnear = b - sqrt(disc);

  if (tnear < 0.0)
    discard;

  // calculate hit point and resulting surface normal
  vec3 pnt = rayorigin + tnear * raydir;
  vec3 N = normalize(pnt - spherepos);

  // Output the ray-sphere intersection point as the fragment depth
  // rather than the depth of the bounding box polygons.
  // The eye coordinate Z value must be transformed to normalized device
  // coordinates before being assigned as the final fragment depth.
  if (vmdprojectionmode == 1) {
    // perspective projection = 0.5 + (hfpn + (f * n / pnt.z)) / diff
    gl_FragDepth = 0.5 + (vmdprojparms[2] + (vmdprojparms[1] * vmdprojparms[(
3];
  } else {
    // orthographic projection = 0.5 + (-hfpn - pnt.z) / diff
    gl_FragDepth = 0.5 + (-vmdprojparms[2] - pnt.z) / vmdprojparms[3];
  }

#ifdef TEXTURE
  // perform texturing operations for volumetric data
  // The only texturing mode that applies to the sphere shader
```

Use of GPUs For Computation

- Widespread support for programmable shading led researchers to begin experimenting with the use of GPUs for general purpose computation, “GPGPU”
- Early GPGPU efforts used existing graphics APIs to express computation in terms of drawing
- As one would expect, expressing general computation problems in terms of triangles and pixels and “drawing the answer” is obfuscating and painful to debug to say the least…
- Soon researchers began creating dedicated GPU programming tools, starting with Brook and Sh, and ultimately leading to a variety of commercial tools such as CUDA

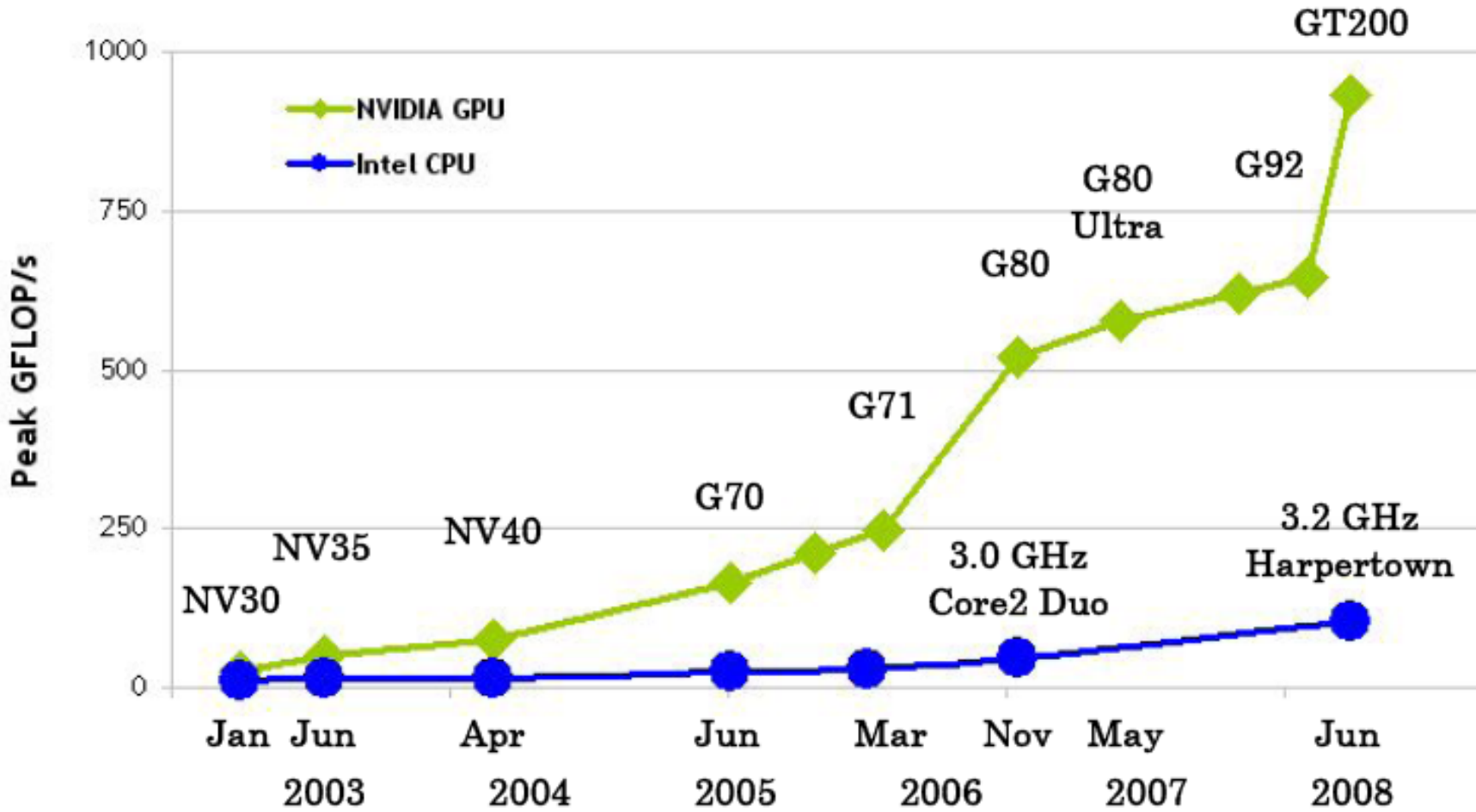
GPU Computing

- Commodity devices, omnipresent in modern computers
- Massively parallel hardware, hundreds of processing units, throughput oriented design
- Support all standard integer and floating point types
- Programming tools allow software to be written in dialects of familiar C/C++ and integrated into legacy software
- GPU algorithms are often multicore-friendly due to attention paid to data locality and work decomposition, and can be successfully executed on multi-core CPUs as well, using special runtime systems (e.g. MCUDA)

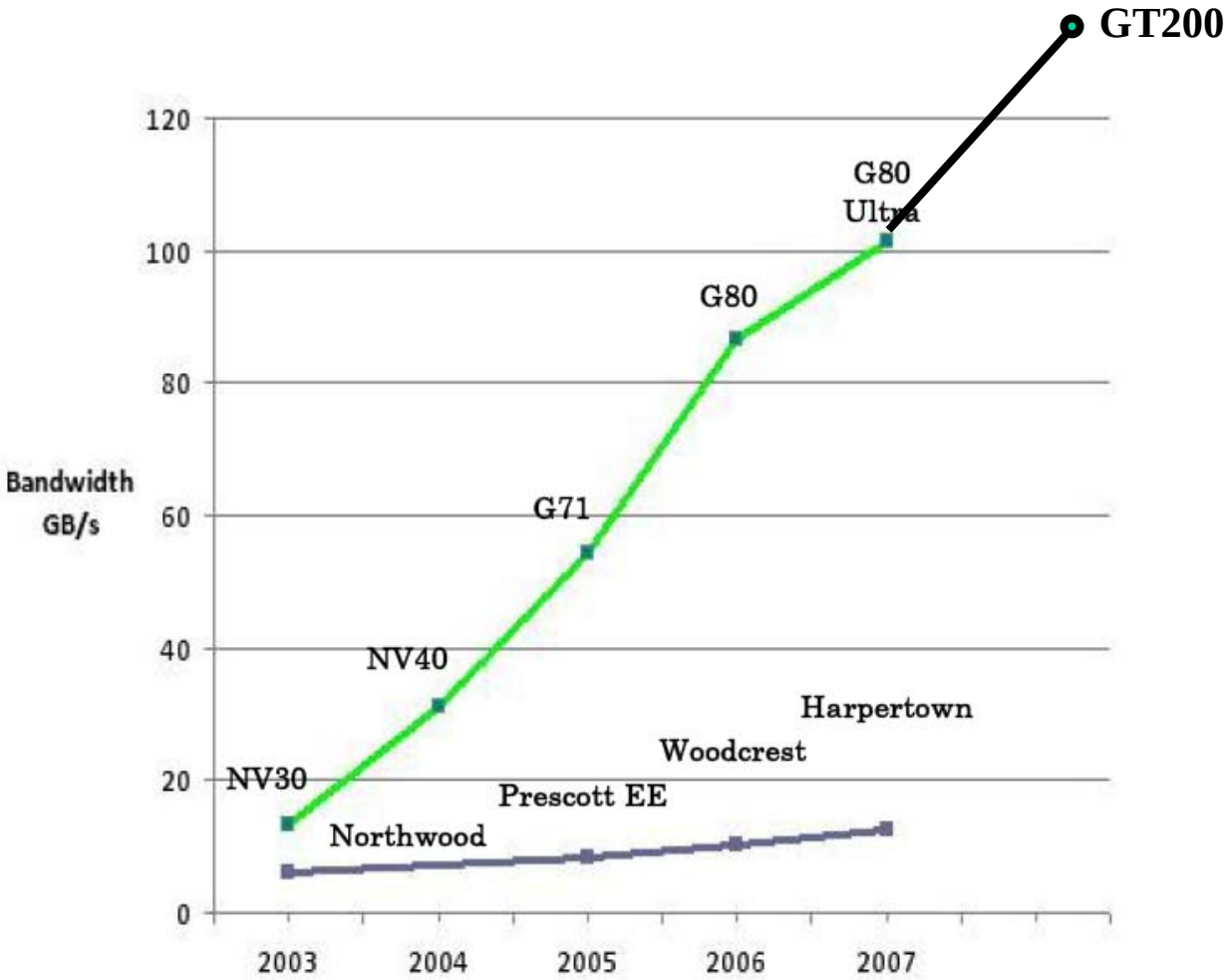
What Speedups Can GPUs Achieve?

- Single-GPU speedups of 8x to 30x vs. CPU core are quite common
- Best speedups (100x!) are attained on codes that are skewed towards floating point arithmetic, esp. CPU-unfriendly operations that prevent effective use of SSE or other vectorization
- Amdahl's Law can prevent legacy codes from achieving peak speedups with only shallow GPU acceleration efforts

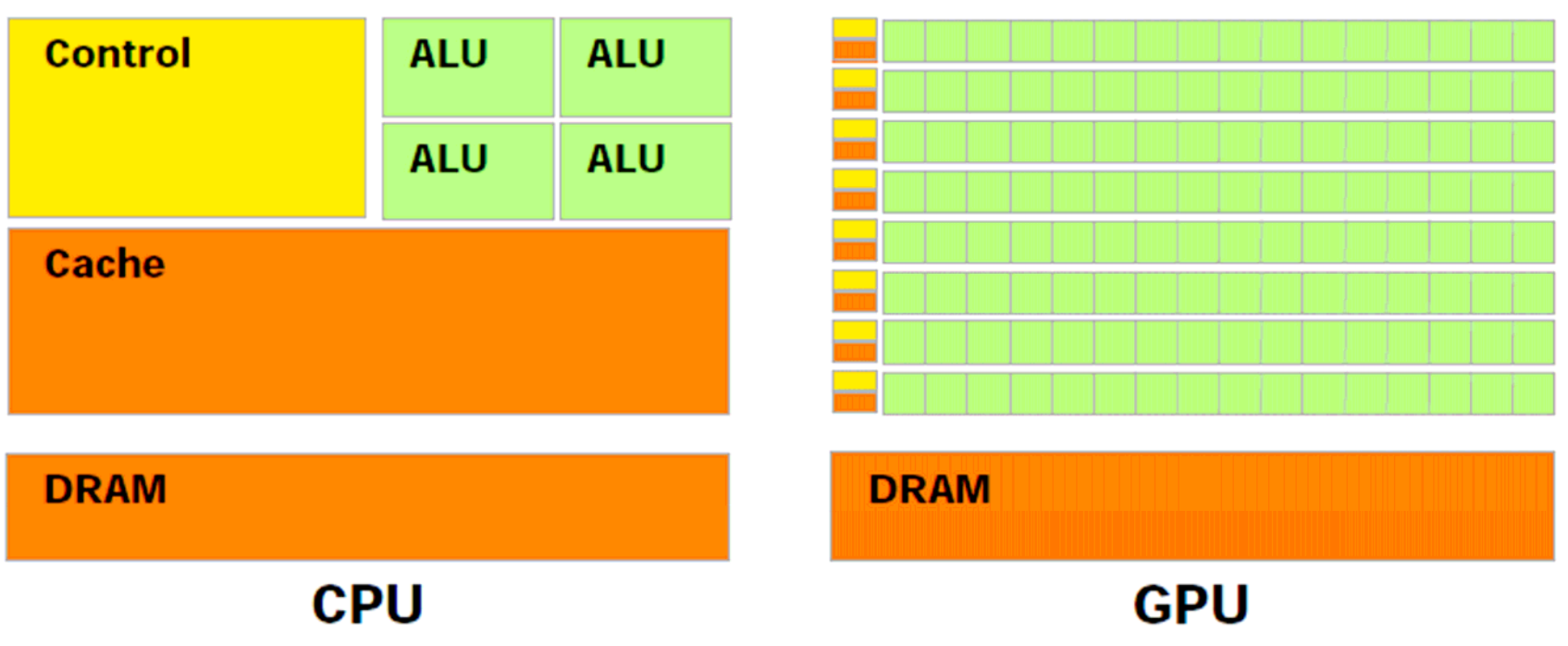
Peak Single-precision Arithmetic Performance Trend

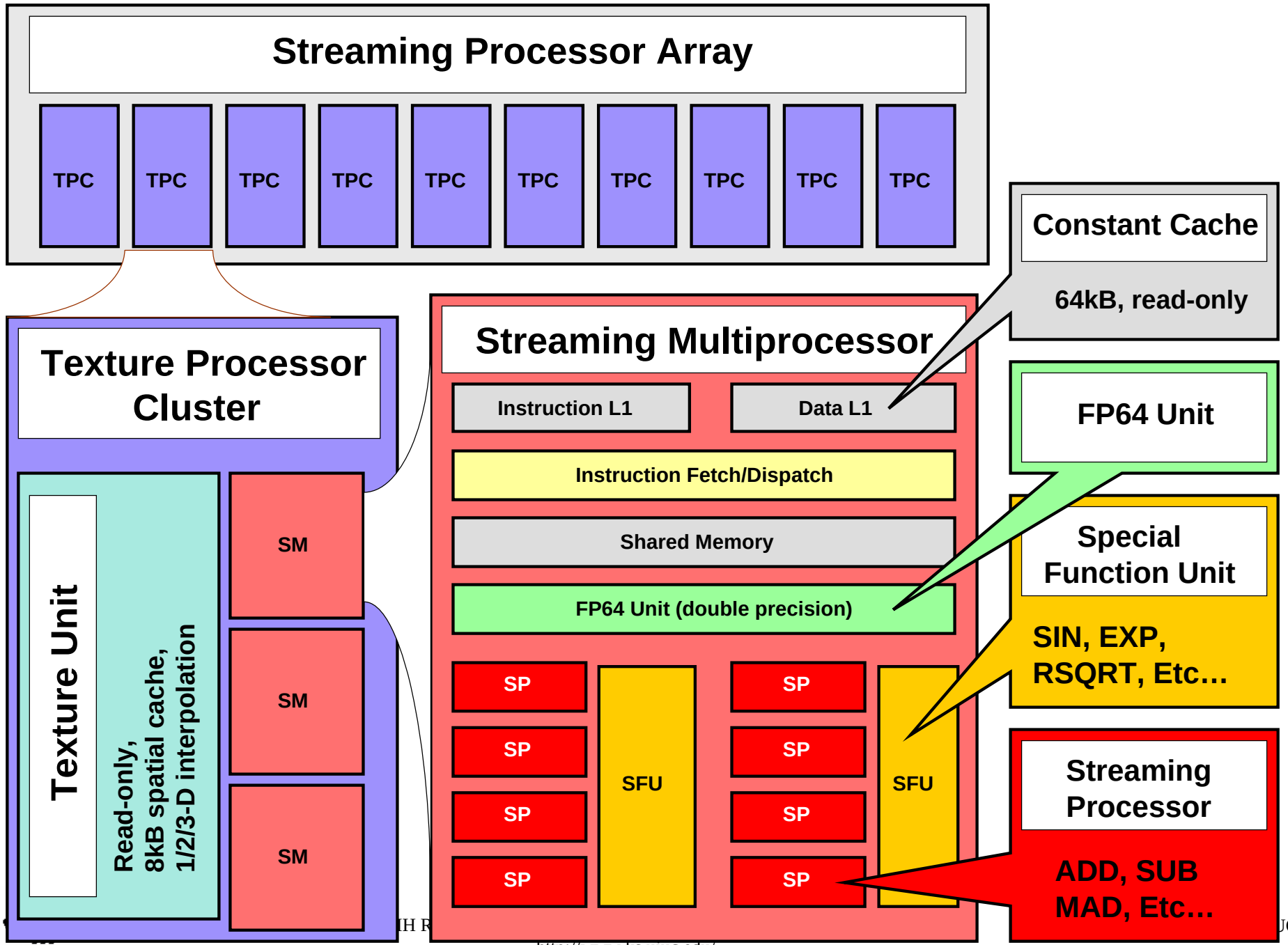


Peak Memory Bandwidth Trend



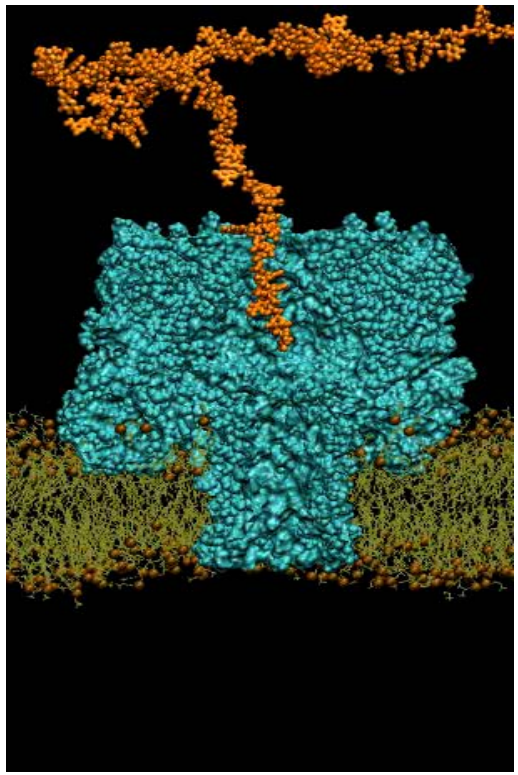
Comparison of CPU and GPU Hardware Architecture




Streaming Processor Array
TPC
TPC
TPC
TPC
TPC
TPC
TPC
TPC
TPC
TPC
Texture Processor Cluster
Texture Unit
Read-only, 8kB spatial cache, 1/2/3-D interpolation
SM
SM
SM
Streaming Multiprocessor
Instruction L1
Data L1
Instruction Fetch/Dispatch
Shared Memory
FP64 Unit (double precision)
SP
SP
SP
SP
SFU
SP
SP
SP
SP
SFU
Constant Cache
64kB, read-only
FP64 Unit
Special Function Unit
SIN, EXP, RSQRT, Etc…
Streaming Processor
ADD, SUB MAD, Etc…

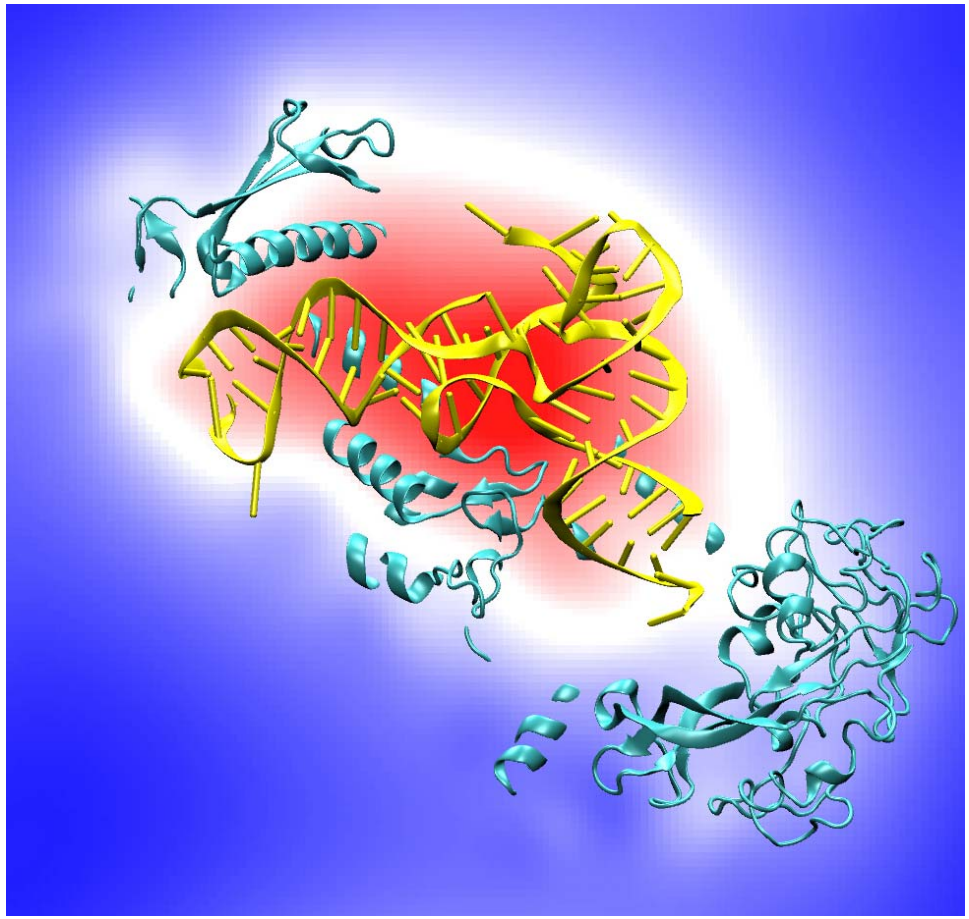
Computational Biology's Insatiable Demand for Processing Power

- Simulations still fall short of biological timescales
- Large simulations extremely difficult to prepare, analyze
- Order of magnitude increase in performance would allow use of more sophisticated models



Calculating Electrostatic Potential Maps

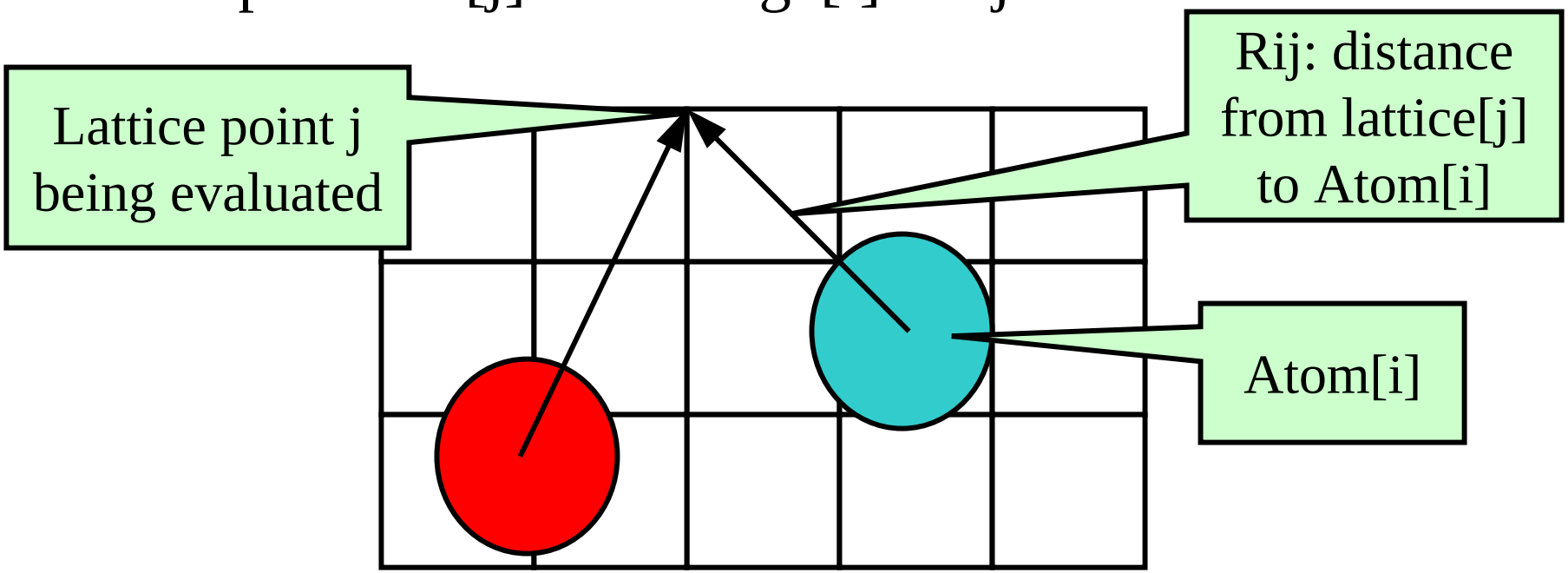
- Used in molecular structure building, analysis, visualization, simulation
- Electrostatic potentials evaluated on a uniformly spaced 3-D lattice
- Each lattice point contains sum of electrostatic contributions of all atoms



Direct Coulomb Summation

- At each lattice point, sum potential contributions for all atoms in the simulated structure:

potential[j] += charge[i] / Rij



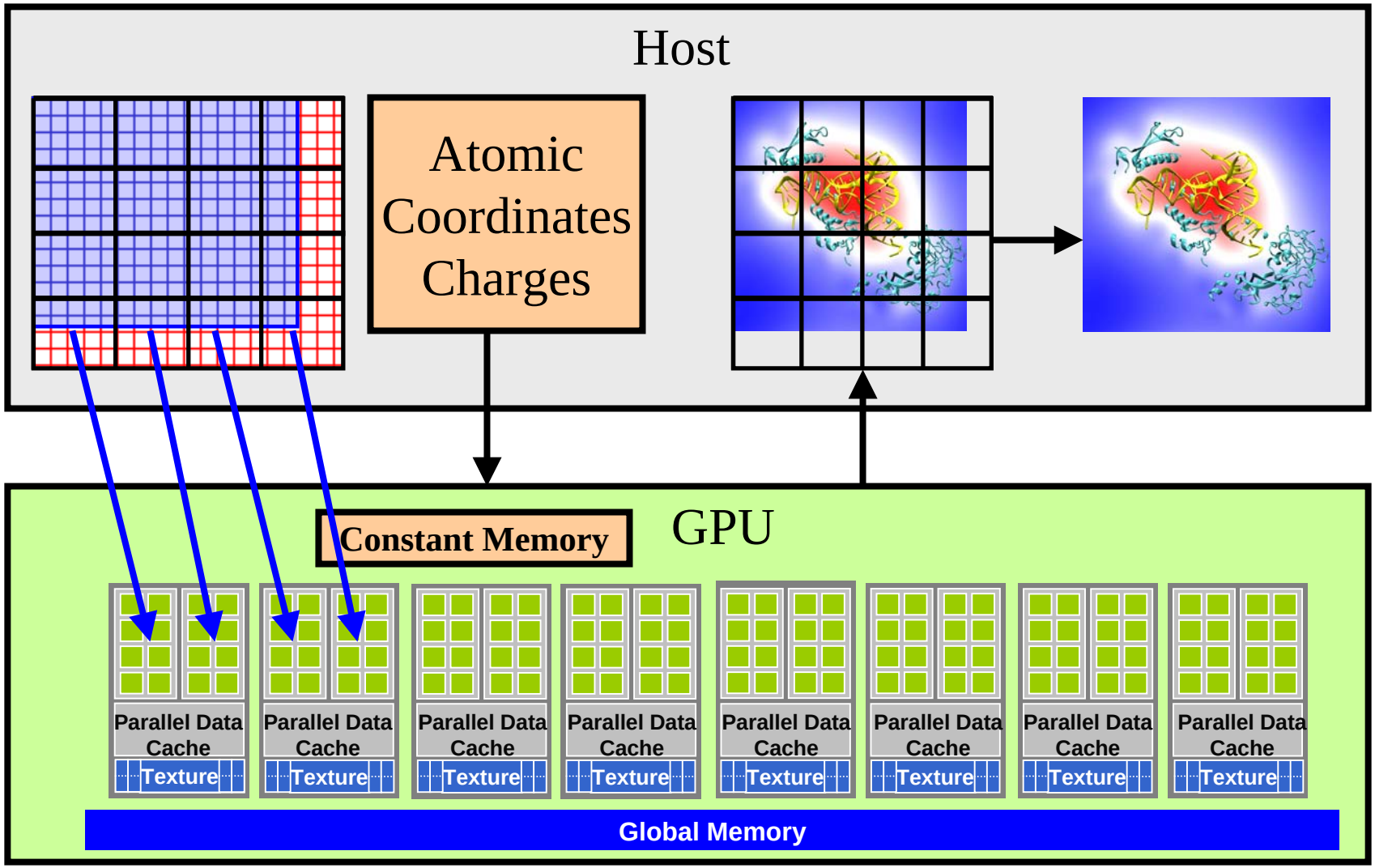
Single Slice DCS: Simple (Slow) C Version

```
void cenergy(float *energygrid, dim3 grid, float gridspacing, float z, const float *atoms, int numatoms) {
  int i,j,n;
  int atomarrdim = numatoms * 4;
  for (j=0; j<grid.y; j++) {
    float y = gridspacing * (float) j;
    for (i=0; i<grid.x; i++) {
      float x = gridspacing * (float) i;
      float energy = 0.0f;
      for (n=0; n<atomarrdim; n+=4) {     // calculate potential contribution of each atom
        float dx = x - atoms[n    ];
        float dy = y - atoms[n+1];
        float dz = z - atoms[n+2];
        energy += atoms[n+3] / sqrtf(dx*dx + dy*dy + dz*dz);
      }
      energygrid[grid.x*grid.y*k + grid.x*j + i] = energy;
    }
  }
```

Direct Coulomb Summation on the GPU

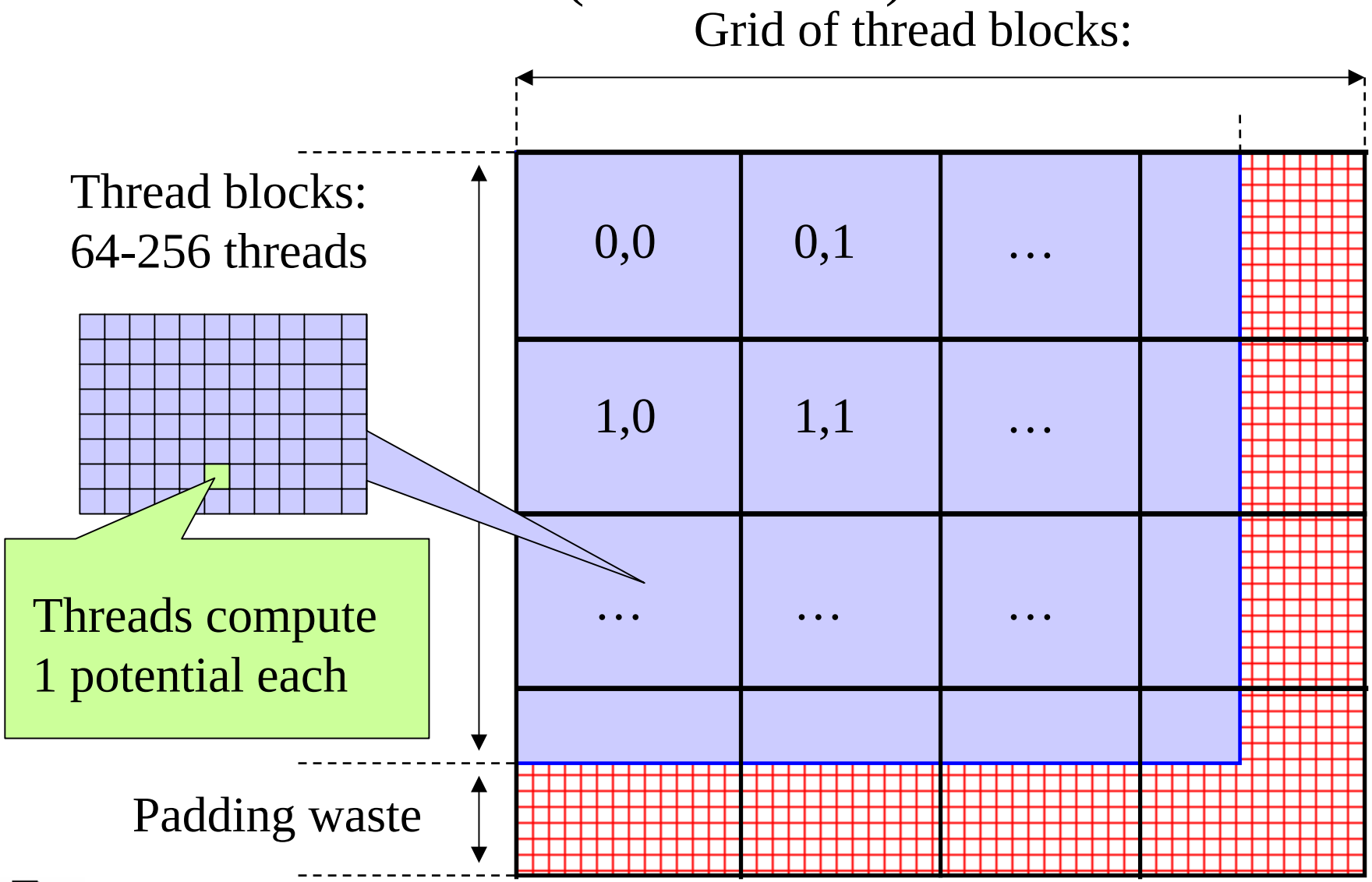
- GPU outruns a CPU core by 44x
- Work is decomposed into tens of thousands of independent threads, multiplexed onto hundreds of GPU processing units
- Single-precision FP arithmetic is adequate for intended application
- Numerical accuracy can be further improved by compensated summation, spatially ordered summation groupings, or accumulation of potential in double-precision
- Starting point for more sophisticated algorithms

Direct Coulomb Summation on the GPU



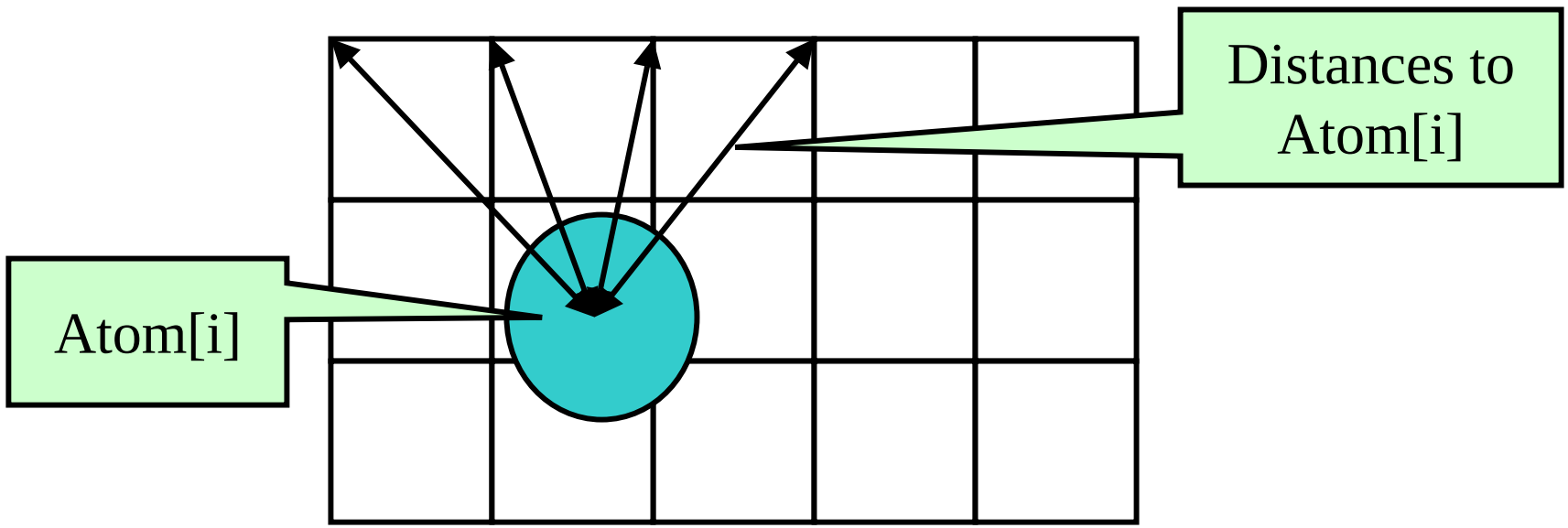
DCS CUDA Block/Grid Decomposition

(non-unrolled)



DCS CUDA Algorithm: Unrolling Loops

- Reuse atom data and partial distance components multiple times
- Add each atom's contribution to several lattice points at a time, where distances only differ in one component

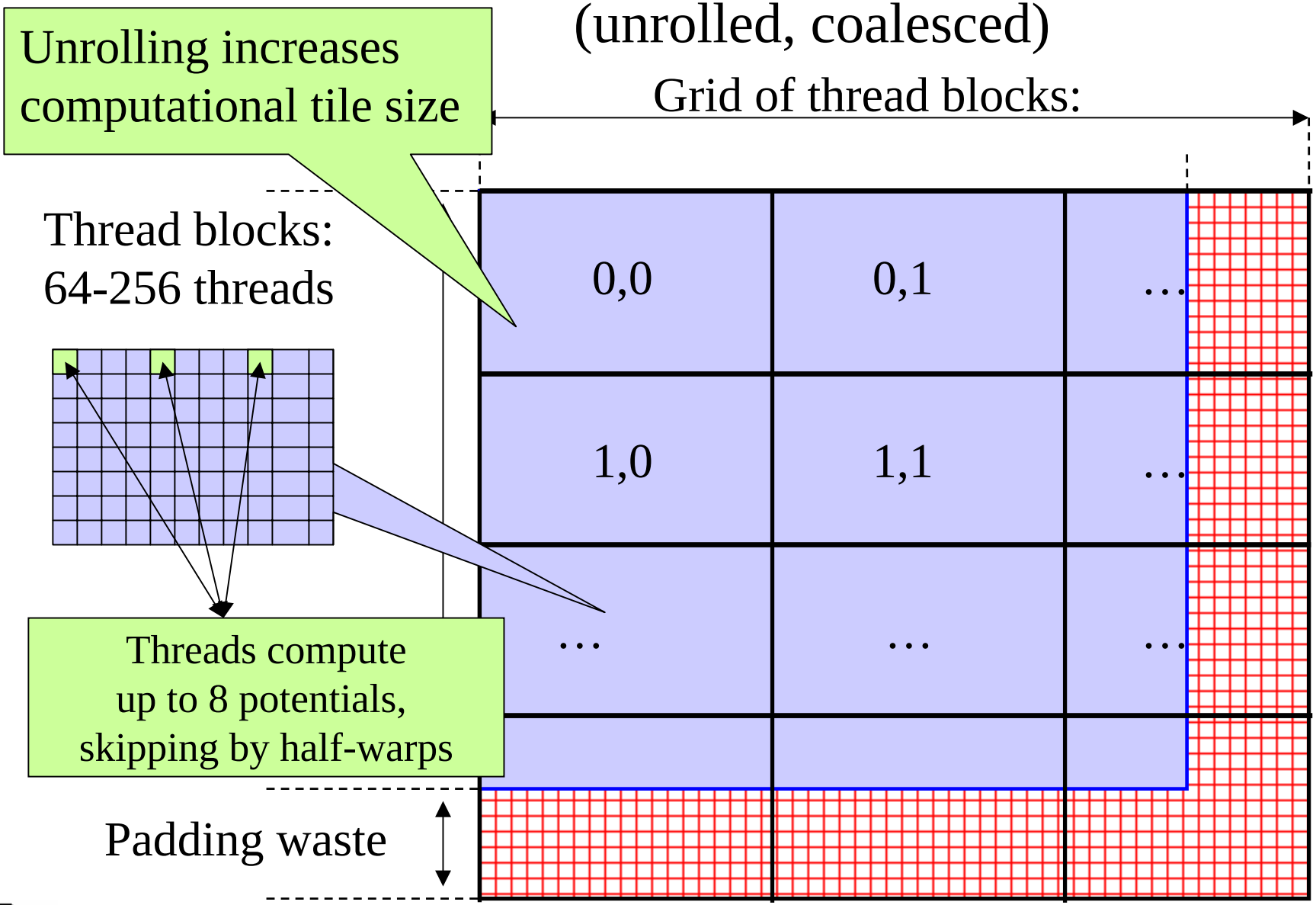


DCS Inner Loop (Unroll and Jam)

```
…
  for (atomid=0; atomid<numatoms; atomid++) {
    float dy = coory - atominfo[atomid].y;
    float dysqpdzsq = (dy * dy) + atominfo[atomid].z;
    float dx1 = coorx1 - atominfo[atomid].x;
    float dx2 = coorx2 - atominfo[atomid].x;
    float dx3 = coorx3 - atominfo[atomid].x;
    float dx4 = coorx4 - atominfo[atomid].x;
    energyvalx1 += atominfo[atomid].w * rsqrtf(dx1*dx1 + dysqpdzsq);
    energyvalx2 += atominfo[atomid].w * rsqrtf(dx2*dx2 + dysqpdzsq);
    energyvalx3 += atominfo[atomid].w * rsqrtf(dx3*dx3 + dysqpdzsq);
    energyvalx4 += atominfo[atomid].w * rsqrtf(dx4*dx4 + dysqpdzsq);
  }
…
```

DCS CUDA Block/Grid Decomposition

(unrolled, coalesced)

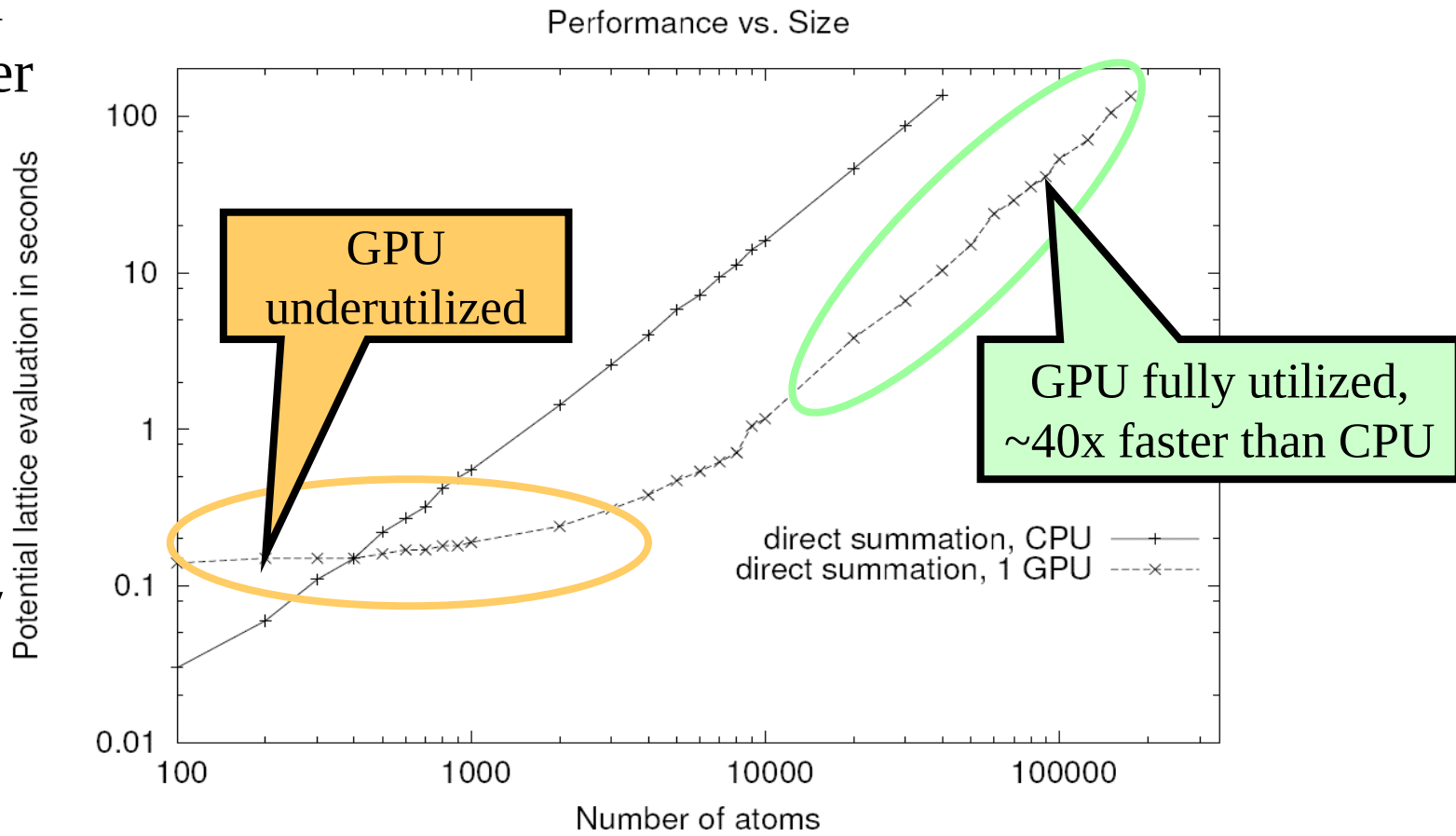


DCS Version 4: Kernel Structure
291.5 GFLOPS, 39.5 Billion Atom Evals/Sec

- Processes 8 lattice points at a time in the inner loop
- Subsequent lattice points computed by each thread are offset to guarantee coalesced memory accesses
- Loads and increments 8 potential map lattice points from global memory at completion of of the summation, avoiding register consumption
- Code is too long to show, but is available by request

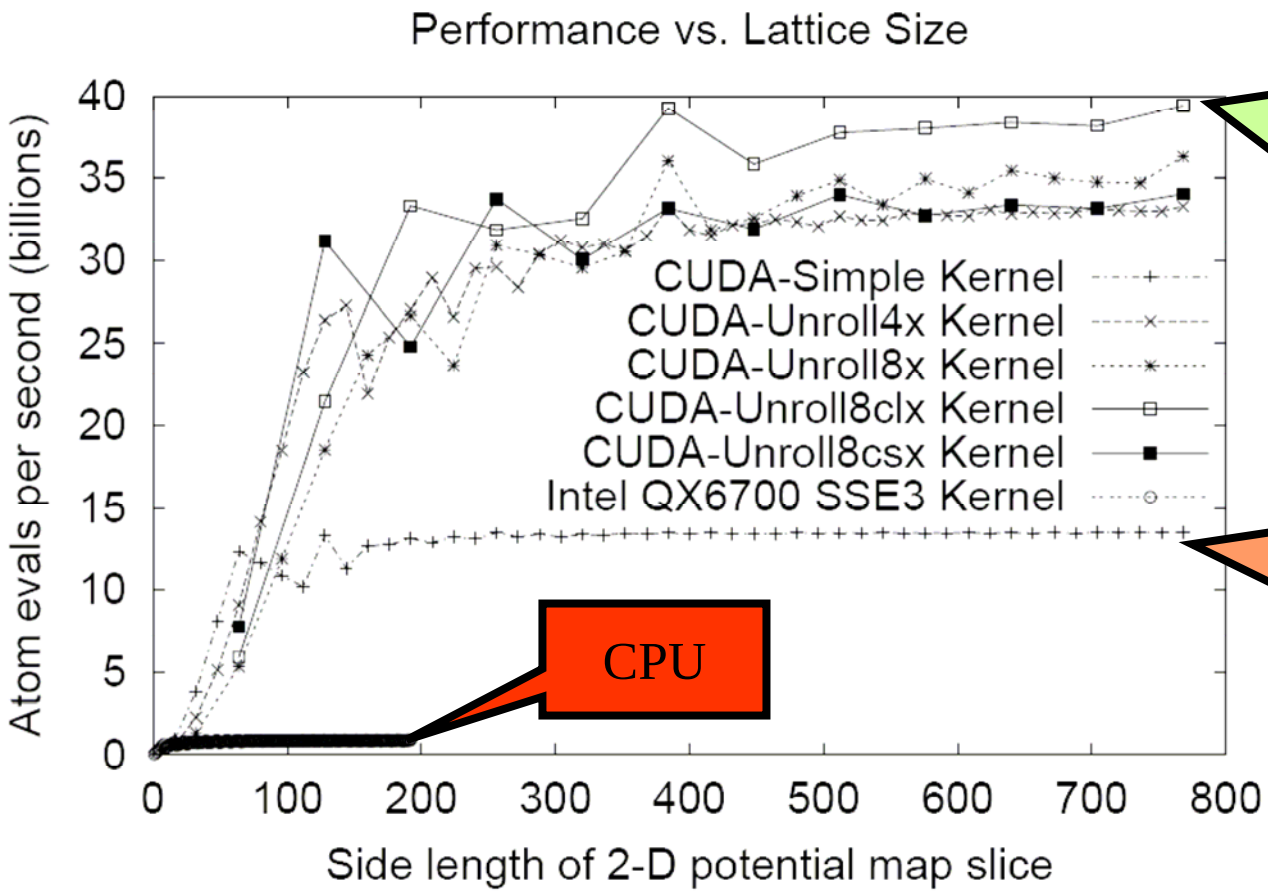
Direct Coulomb Summation Runtime

Lower is better



Accelerating molecular modeling applications with graphics processors. J. Stone, J. Phillips, P. Freddolino, D. Hardy, L. Trabuco, K. Schulten. *J. Comp. Chem.*, 28:2618-2640, 2007.

Direct Coulomb Summation Performance



CUDA-Unroll8clx: fastest GPU kernel, 44x faster than CPU, 291 GFLOPS on GeForce 8800GTX

CUDA-Simple: 14.8x faster, 33% of fastest GPU kernel

GPU computing. J. Owens, M. Houston, D. Luebke, S. Green, J. Stone, J. Phillips. *Proceedings of the IEEE*, 96:879-899, 2008.

Multi-GPU Direct Coulomb Summation

- 4-GPU (2 Quadroplex) Opteron node at NCSA
- 157 billion evals/sec
- 1.16 TFLOPS
- 176x speedup vs. Intel QX6700 CPU core w/ SSE

- 4-GPU GTX 280 (GT200)
- 241 billion evals/sec
- 1.78 TFLOPS
- 271x speedup vs. Intel QX6700 CPU core w/ SSE

NCSA GPU Cluster

http://www.ncsa.uiuc.edu/Projects/GPUcluster/

Infinite vs. Cutoff Potentials

- Infinite range potential:
  - All atoms contribute to all lattice points
  - Summation algorithm has quadratic complexity
- Cutoff (range-limited) potential:
  - Atoms contribute within cutoff distance to lattice points
  - Summation algorithm has linear time complexity
  - Has many applications in molecular modeling:
    - Replace electrostatic potential with shifted form
    - Short-range part for fast methods of approximating full electrostatics
    - Used for fast decaying interactions (e.g. Lennard-Jones, Buckingham)

NAMD Parallel Molecular Dynamics

Kale et al., J. Comp. Phys. 151:283-312, 1999.

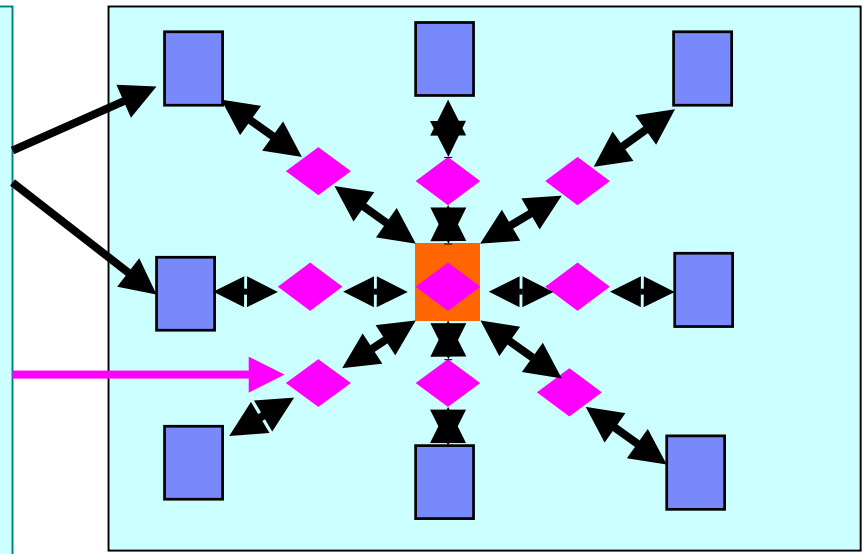
- Designed from the beginning as a parallel program
- Uses the Charm++ philosophy:
  - Decompose computation into a large number of objects
  - Intelligent Run-time system (Charm++) assigns objects to processors for dynamic load balancing with minimal communication

Hybrid of spatial and force decomposition:

•Spatial decomposition of atoms into cubes (called patches)

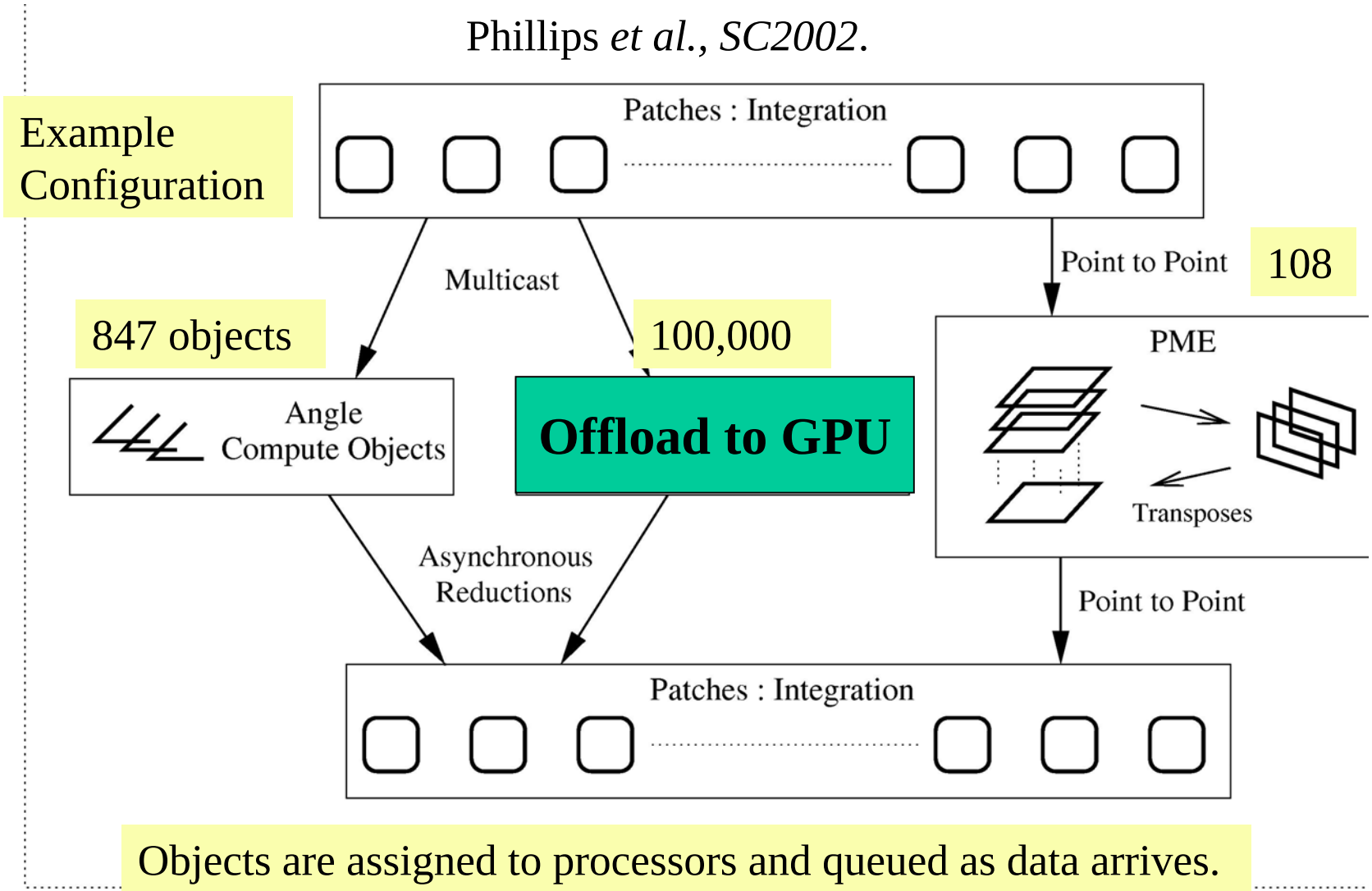
•For every pair of interacting patches, create one object for calculating electrostatic interactions

•Recent: Blue Matter, Desmond, etc. use this idea in some form



NAMD Overlapping Execution

Phillips *et al.*, *SC2002*.



Nonbonded Forces CUDA Code

```
texture<float4> force_table;
__constant__ unsigned int exclusions[];
__shared__ atom jatom[];
atom iatom;      // per-thread atom, stored in registers
float4 iforce;   // per-thread force, stored in registers
for ( int j = 0; j < jatom_count; ++j ) {
  float dx = jatom[j].x - iatom.x;   float dy = jatom[j].y - iatom.y;  float dz = jatom[j].z - iatom.z;
  float r2 = dx*dx + dy*dy + dz*dz;
  if ( r2 < cutoff2 ) {
    float4 ft = texfetch(force_table, 1.f/sqrt(r2));
    bool excluded = false;
    int indexdiff = iatom.index - jatom[j].index;
    if ( abs(indexdiff) <= (int) jatom[j].excl_maxdiff ) {
      indexdiff += jatom[j].excl_index;
      excluded = ((exclusions[indexdiff>>5] & (1<<(indexdiff&31))) != 0);
    }
    float f = iatom.half_sigma + jatom[j].half_sigma;  // sigma
    f *= f*f;  // sigma^3
    f *= f;  // sigma^6
    f *= ( f * ft.x + ft.y );  // sigma^12 * fi.x - sigma^6 * fi.y
    f *= iatom.sqrt_epsilon * jatom[j].sqrt_epsilon;
    float qq = iatom.charge * jatom[j].charge;
    if ( excluded ) { f = qq * ft.w; }  // PME correction
    else { f += qq * ft.z; }  // Coulomb
    iforce.x += dx * f;   iforce.y += dy * f;    iforce.z += dz * f;
    iforce.w += 1.f;  // interaction count or energy
  }
}
```

Force Interpolation

Exclusions

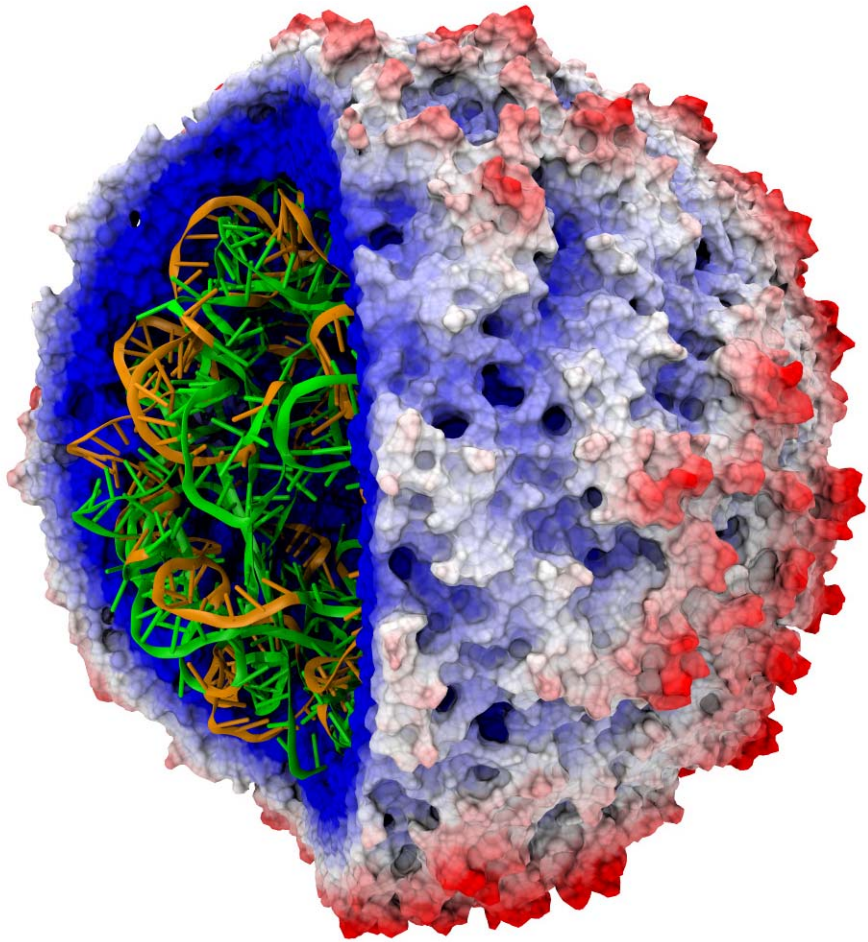
Parameters

Accumulation

Stone *et al.*, *J. Comp. Chem.* **28**:2618-2640, 2007.

Molecular Simulations: Virology

- Simulations lead to better understanding of the mechanics of viral infections
- Better understanding of infection mechanics at the molecular level may result in more effective treatments for diseases
- Since viruses are large, their computational “viewing” requires tremendous resources, in particular large parallel computers
- GPUs can significantly accelerate the simulation, analyses, and visualization of such structures



Satellite Tobacco Mosaic Virus (STMV)

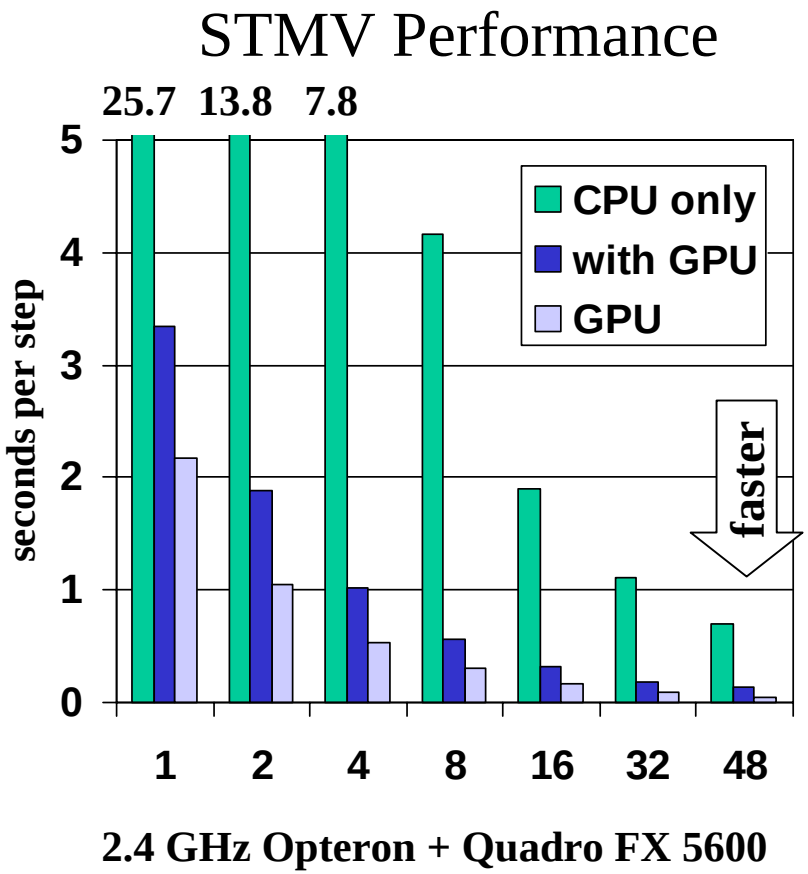
NAMD Performance on NCSA GPU Cluster, April 2008

CPU Cores & GPUs	4	8	16	32	60
GPU-accelerated performance					
Local blocks/GPU	13186	5798	2564	1174	577
Remote blocks/GPU	1644	1617	1144	680	411
GPU s/step	0.544	0.274	0.139	0.071	0.040
Total s/step	0.960	0.483	0.261	0.154	0.085
Unaccelerated performance					
Total s/step	6.76	3.33	1.737	0.980	0.471
Speedup from GPU acceleration					
Factor	7.0	6.9	6.7	6.4	5.5

STMV benchmark, 1M atoms,12A cutoff, PME every 4 steps, on 2.4 GHz AMD Opteron + NVIDIA Quadro FX 5600

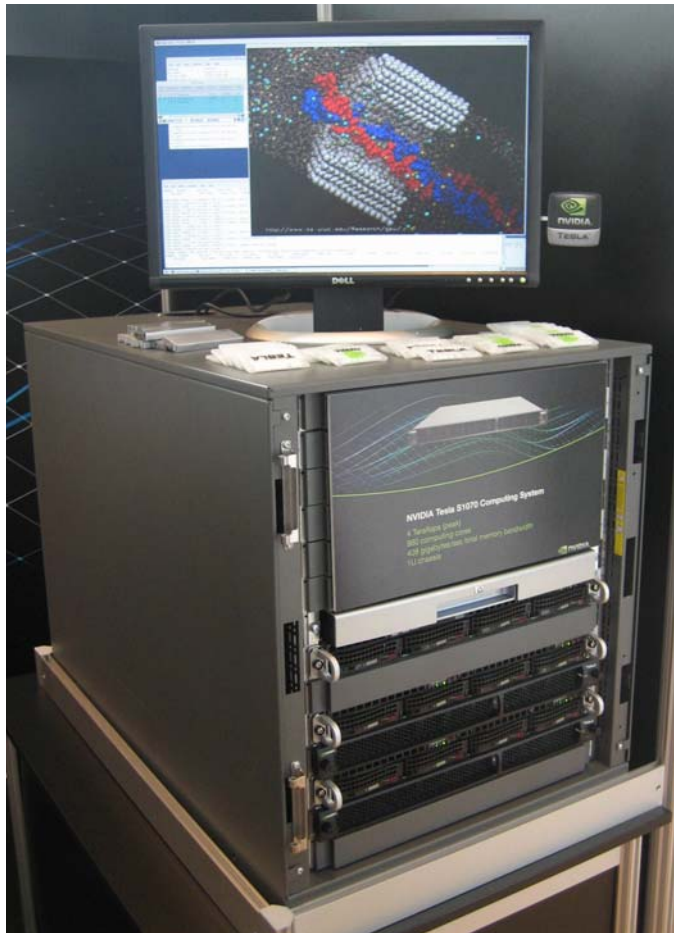
NAMD Performance on NCSA GPU Cluster, April 2008

- STMV virus (1M atoms)
- 60 GPUs match performance of 330 CPU cores
- 5.5-7x overall application speedup w/ G80-based GPUs
- Overlap with CPU
- Off-node results done first
- Plans for better performance
  - Tune or port remaining work
  - Balance GPU load



NAMD Performance on GT200 GPU Cluster, August 2008

- 8 GT200s, 240 SPs @ 1.3GHz:
  - 72x faster than a single CPU core
  - 9x overall application speedup vs. 8 CPU cores
  - 32% faster overall than 8 nodes of G80 cluster
  - GT200 CUDA kernel is 54% faster
  - ~8% variation in GPU load
- Cost of double-precision for force accumulation is minimal: only 8% slower than single-precision



GPU Kernel Performance, May 2008

GeForce 8800GTX w/ CUDA 1.1, Driver 169.09

http://www.ks.uiuc.edu/Research/gpu/

Calculation / Algorithm	Algorithm class	Speedup vs. Intel QX6700 CPU core
Fluorescence microphotolysis	Iterative matrix / stencil	12x
Pairlist calculation	Particle pair distance test	10-11x
Pairlist update	Particle pair distance test	5-15x
Molecular dynamics non-bonded force calc.	N-body cutoff force calculations	10x 20x (w/ pairlist)
Cutoff electron density sum	Particle-grid w/ cutoff	15-23x
MSM short-range	Particle-grid w/ cutoff	24x
MSM long-range	Grid-grid w/ cutoff	22x
Direct Coulomb summation	Particle-grid	44x

Lessons Learned

- GPU algorithms need fine-grained parallelism and sufficient work to fully utilize the hardware
- Fine-grained GPU work decompositions compose well with the comparatively coarse-grained decompositions used for multicore or distributed memory programing
- Much of GPU algorithm optimization revolves around efficient use of multiple memory systems and latency hiding

Lessons Learned (2)

- The host CPU can potentially be used to “regularize” the computation for the GPU, yielding better overall performance
- Overlapping CPU work with GPU can hide some communication and unaccelerated computation

Acknowledgement

- Additional Information and References:
  - **http://www.ks.uiuc.edu/Research/gpu/**
- Acknowledgement, questions, source code requests:
  - John Stone (johns@ks.uiuc.edu)
  - Theoretical and Computational Biophysics Group, NIH Resource for Macromolecular Modeling and Bioinformatics
    Beckman Institute for Advanced Science and Technology,
  - NCSA GPU Cluster
  - NVIDIA
- NIH funding: P41-RR05969

Publications

http://www.ks.uiuc.edu/Research/gpu/

- Adapting a message-driven parallel application to GPU-accelerated clusters. J. Phillips, J. Stone, K. Schulten. *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*, *(in press)*
- GPU acceleration of cutoff pair potentials for molecular modeling applications. C. Rodrigues, D. Hardy, J. Stone, K. Schulten, W. Hwu. *Proceedings of the 2008 Conference On Computing Frontiers*, pp. 273-282, 2008.
- GPU computing. J. Owens, M. Houston, D. Luebke, S. Green, J. Stone, J. Phillips. *Proceedings of the IEEE*, 96:879-899, 2008.
- Accelerating molecular modeling applications with graphics processors. J. Stone, J. Phillips, P. Freddolino, D. Hardy, L. Trabuco, K. Schulten. *J. Comp. Chem.*, 28:2618-2640, 2007.
- Continuous fluorescence microphotolysis and correlation spectroscopy. A. Arkhipov, J. Hüve, M. Kahms, R. Peters, K. Schulten. *Biophysical Journal*, 93:4006-4017, 2007.