Improving performance of the N-Body problem

Efim Sergeev

Senior Software Engineer at Singularis Lab LLC
Contents

• Theory
• Naive version
• Memory layout optimization
• Cache Blocking Techniques
• Data Alignment to Assist Vectorization
Force Calculation: Direct Summation

- Given $N$ bodies with initial position $\mathbf{x}_i$ and velocity $\mathbf{v}_i$, the force vector $\mathbf{f}_{ij}$ on body $i$ caused by its gravitational attraction to body $i$ is given by following:

$$ \mathbf{f}_{ij} = G \frac{m_i m_j}{\left\| \mathbf{r}_{ij} \right\|^2} \cdot \frac{\mathbf{r}_{ij}}{\left\| \mathbf{r}_{ij} \right\|}$$

- where $m_i$ and $m_j$ are masses of bodies;
- $\mathbf{r}_{ij} = \mathbf{x}_j - \mathbf{x}_i$ is the vector from body $i$ to $j$;
- $G$ is the gravitational constant.
Force Calculation: Direct Summation

- The total force $F_i$ on body $i$, due to its interactions with the other $N-1$ bodies, is obtained by summing all interactions:

$$F_i = \sum_{1 \leq j \leq N, j \neq i} f_{ij} = G m_i \cdot \sum_{1 \leq i \leq N, j \neq i} \frac{m_i r_{ij}}{||r_{ij}||^3}$$

- **Advantages**: robust, accurate, completely general.
- **Disadvantage**: computational cost per body is $O(N)$, need $O(N^2)$ operations to compute forces on all bodies.
- **However, direct summation is a good fit with**:
  1. Individual time steps;
  2. Specialized hardware;
typedef struct { float x, y, z, vx, vy, vz; } Body;

void bodyForce(Body *p, float dt, int n) {
    #pragma omp parallel for schedule(dynamic)
    for (int i = 0; i < n; i++) {
        float Fx = 0.0f; float Fy = 0.0f; float Fz = 0.0f;

        for (int j = 0; j < n; j++) {
            float dx = p[j].x - p[i].x;
            float dy = p[j].y - p[i].y;
            float dz = p[j].z - p[i].z;
            float distSqr = dx*dx + dy*dy + dz*dz + SOFTENING;
            float invDist = 1.0f / sqrtf(distSqr);
            float invDist3 = invDist * invDist * invDist;

            Fx += dx * invDist3; Fy += dy * invDist3; Fz += dz * invDist3;
        }

        p[i].vx += dt*Fx; p[i].vy += dt*Fy; p[i].vz += dt*Fz;
    }
}
Naive version compile & run

- Allocate resources
  - `#salloc -N 1 --partition=mic`
  - `#squeue`
- Connect to node
  - `#ssh %nodename%`
  - `#cd ./Workshop2016/src/mini-nbody`
- Compile
  - `#./build-nbody-naive.sh`
- Run
  - `# ./run-nbody-naive.sh`

**Interaction per second**

<table>
<thead>
<tr>
<th>Number of bodies</th>
<th>Interaction per second ($10^9$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td>0</td>
</tr>
<tr>
<td>2048</td>
<td>0</td>
</tr>
<tr>
<td>4069</td>
<td>0</td>
</tr>
<tr>
<td>8192</td>
<td>0</td>
</tr>
<tr>
<td>16384</td>
<td>10.5</td>
</tr>
<tr>
<td>32768</td>
<td>15.0</td>
</tr>
<tr>
<td>65536</td>
<td>10.0</td>
</tr>
<tr>
<td>131072</td>
<td>15.0</td>
</tr>
<tr>
<td>262144</td>
<td>10.0</td>
</tr>
<tr>
<td>524288</td>
<td>15.0</td>
</tr>
<tr>
<td>1048576</td>
<td>10.0</td>
</tr>
</tbody>
</table>
• 60(61) cores
• 4 HW threads/core
• 8 memory controllers 16 Channel GDDR5
• High-speed bi-directional ring interconnect Fully Coherent L2 Cache
Memory Hierarchy for Intel® Xeon Phi™ Core Architecture

- 4 Threads per Core
- 64-bit
- 512-wide
- VPU: Integer, SP, DP, 3-operand, 16-instruction
- 32KB per core
- 512 per core
- Fully Coherent
Memory Hierarchy for Intel® Xeon Phi™

- CPU core and Vector registers
  - 32x64 bytes, 1 cycle
  - 512 avx register, 1 cycle

- L1 Data Cache
- L1 Instruction Cache
  - 32 KB, 3 cycles
  - 32 KB, 3 cycles

- L2 Cache
- Interconnected
  - 512 KB, 11 cycles

- Main Memory
  - 8-16GB, 100+ cycles
Array of structures:

```c
struct {
    float x, y, z, vx, vy, vz;
} Body;
```

Array of structures in RAM looks something like following:

| x0 | y0 | z0 | vx0 | vy0 | vz0 | x1 | y1 | z1 | vx1 |

Disadvantages:

- Use memory gather-scatter operations in vectorized code
- Performing SIMD operations on the original AoS format can require more calculations
- Difficult optimize by compiler
Memory Layout Transformations: Structures of Arrays (SoA)

Structure of array:

```c
struct {
} BodySystem;
```

Array of structures in RAM looks something like following:

<table>
<thead>
<tr>
<th></th>
<th>x0</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>y0</td>
<td>y1</td>
<td>y2</td>
<td>y3</td>
<td>y4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>z0</td>
<td>z1</td>
<td>z2</td>
<td>z3</td>
<td>z4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Advantages:

- Keep memory accesses contiguous when vectorization is performed
- SoA arrangement allows more efficient use of the parallelism of the SIMD technologies because the data is ready for computation in a more optimal vertical manner
- More efficient use of caches and bandwidth
Memory Layout Transformations: Loading in AVX(registers)

Memory: array of structure

<table>
<thead>
<tr>
<th>x0</th>
<th>y0</th>
<th>z0</th>
<th>vx0</th>
<th>vy0</th>
<th>vz0</th>
<th>x1</th>
<th>y1</th>
<th>z1</th>
<th>vx1</th>
</tr>
</thead>
</table>

AVX/SIMD registers

<table>
<thead>
<tr>
<th>x0</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x3</th>
</tr>
</thead>
<tbody>
<tr>
<td>y0</td>
<td>y1</td>
<td>y2</td>
<td>y3</td>
<td>y4</td>
</tr>
<tr>
<td>z0</td>
<td>z1</td>
<td>z2</td>
<td>z3</td>
<td>z4</td>
</tr>
</tbody>
</table>

Memory: structure of array

<table>
<thead>
<tr>
<th>x0</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x3</th>
</tr>
</thead>
<tbody>
<tr>
<td>y0</td>
<td>y1</td>
<td>y2</td>
<td>y3</td>
<td>y4</td>
</tr>
<tr>
<td>z0</td>
<td>z1</td>
<td>z2</td>
<td>z3</td>
<td>z4</td>
</tr>
</tbody>
</table>

AVX/SIMD registers

<table>
<thead>
<tr>
<th>x0</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x3</th>
</tr>
</thead>
<tbody>
<tr>
<td>y0</td>
<td>y1</td>
<td>y2</td>
<td>y3</td>
<td>y4</td>
</tr>
<tr>
<td>z0</td>
<td>z1</td>
<td>z2</td>
<td>z3</td>
<td>z4</td>
</tr>
</tbody>
</table>

void bodyForce(BodySystem p, float dt, int n) {
    #pragma omp parallel for schedule(dynamic)
    for (int i = 0; i < n; i++) {
        float Fx = 0.0f; float Fy = 0.0f; float Fz = 0.0f;

        for (int j = 0; j < n; j++) {
            float dy = p.y[j] - p.y[i];
            float dz = p.z[j] - p.z[i];
            float dx = p.x[j] - p.x[i];
            float distSqr = dx*dx + dy*dy + dz*dz + SOFTENING;
            float invDist = 1.0f / sqrtf(distSqr);
            float invDist3 = invDist * invDist * invDist;

            Fx += dx * invDist3; Fy += dy * invDist3; Fz += dz * invDist3;
        }

        p.vx[i] += dt*Fx; p.vy[i] += dt*Fy; p.vz[i] += dt*Fz;
    }
}
SoA version compile & run

- **Compile**
  - `#./build-nbody-soa.sh`

- **Run**
  - `#./run-nbody-soa.sh`
SoA optimization results

Interaction per second

10^9

Number of bodies

1024  2048  4069  8192  16384  32768  65536  131072  262144  524288  1048576

SoA

naive

SoA
Cache Blocking Techniques

for (int i = 0; i < n; i++) {
    for (int j = 0; j < n; j++) {
        float dy = p.y[j] - p.y[i];
        float dz = p.z[j] - p.z[i];
        float dx = p.x[j] - p.x[i];
    }
}

Cache reloaded for each i body after every tile iteration by j
Cache Blocking Techniques:
Fit bodies in the cache

- L2 cache size 512 KB
- Each body requires in memory 24 bytes
  - 6 fields by 4 byte, total 24 bytes
- How many bodies can be placed in L2 cache?
  - 512000 / 24 = 21300 bodies
Cache Blocking Techniques:
Fit bodies in the cache

Original source:
```c
for (body1 = 0; body1 < NBODIES; body1++) {
    for (body2 = 0; body2 < NBODIES; body2++) {
        OUT[body1] += compute(body1, body2);
    }
}
```

Modified Source (with 1-D blocking):
```c
for (body2 = 0; body2 < NBODIES; body2 += BLOCK) {
    for (body1 = 0; body1 < NBODIES; body1++) {
        for (body22 = 0; body22 < BLOCK; body22++) {
            OUT[body1] += compute(body1, body2 + body22);
        }
    }
}
```
void bodyForce(BodySystem p, float dt, int n, int tileSize) {
    for (int tile = 0; tile < n; tile += tileSize) {
        int to = tile + tileSize;
        if (to > n) to = n;

        #pragma omp parallel for schedule(dynamic)
        for (int i = 0; i < n; i++) {
            float Fx = 0.0f; float Fy = 0.0f; float Fz = 0.0f;

            for (int j = tile; j < to; j++) {
                float dy = p.y[j] - p.y[i];
                float dz = p.z[j] - p.z[i];
                float dx = p.x[j] - p.x[i];
                float distSqr = dx*dx + dy*dy + dz*dz + SOFTENING;
                float invDist = 1.0f / sqrtf(distSqr);
                float invDist3 = invDist * invDist * invDist;
                Fx += dx * invDist3; Fy += dy * invDist3; Fz += dz * invDist3;
            }

            p.vx[i] += dt*Fx; p.vy[i] += dt*Fy; p.vz[i] += dt*Fz;
        }
    }
}
Cache optimization version compile & run

- **Compile**
  - `./build-nbody-block.sh`

- **Run**
  - `./run-nbody-block.sh`
Cache optimization results

<table>
<thead>
<tr>
<th></th>
<th>Size</th>
<th>Naive</th>
<th>SoA</th>
<th>Cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td></td>
<td>0.424</td>
<td>6.217</td>
<td>6.213</td>
</tr>
<tr>
<td>2048</td>
<td></td>
<td>0.633</td>
<td>12.714</td>
<td>14.191</td>
</tr>
<tr>
<td>4096</td>
<td></td>
<td>1.112</td>
<td>22.297</td>
<td>25.116</td>
</tr>
<tr>
<td>8192</td>
<td></td>
<td>3.387</td>
<td>31.143</td>
<td>35.755</td>
</tr>
<tr>
<td>16384</td>
<td></td>
<td>9.632</td>
<td>25.103</td>
<td>28.303</td>
</tr>
<tr>
<td>32768</td>
<td></td>
<td>9.447</td>
<td>31.228</td>
<td>35.431</td>
</tr>
<tr>
<td>65536</td>
<td></td>
<td>9.248</td>
<td>35.035</td>
<td>42.645</td>
</tr>
<tr>
<td>131072</td>
<td></td>
<td>9.377</td>
<td>33.816</td>
<td>49.333</td>
</tr>
<tr>
<td>262144</td>
<td></td>
<td>9.411</td>
<td>34.239</td>
<td>52.446</td>
</tr>
<tr>
<td>524288</td>
<td></td>
<td>9.478</td>
<td>14.138</td>
<td>52.980</td>
</tr>
<tr>
<td>1048576</td>
<td></td>
<td>9.515</td>
<td>15.6</td>
<td>53.184</td>
</tr>
</tbody>
</table>

Number of bodies

Interaction per second $10^9$
Reason for Vectorization Fails & How to Succeed

• **Alignment:**
  • Align your arrays/data structures

• Most frequent reason is **dependence:**
  • Minimize dependencies among iteration

• Non-unit stride between elements:
  • Possible to change algorithm to allow linear/consecutive access?
Memory alignment

• All memory-based instructions in the coprocessor should be aligned to avoid instruction faults
  – Force the compiler to create data objects with starting addresses that are modulo 64 bytes.
• The compiler is able to make optimizations when the data access (including base-pointer plus index) is known to be aligned by 64 bytes
  – Compiler cannot know nor assume data is aligned inside loops without some help from the user
Memory alignment

L2 cache composed of 64 byte cache lines

64 byte cache lines

RAM read/write alignment

0 bytes
64 bytes
128 bytes
Array in memory not alignment to 64 bytes

L2 CACHE

DRAM

Array in memory loading to cache in 3 operations
Memory alignment

Array in memory not alignment to 64 bytes

64 byte cache lines

every second load will be across a cache-line split

Array in L2 CACHE loading to AVX register in 3 operations
Memory alignment

Array in memory alignment to 64 bytes

L2 CACHE

DRAM

64 byte cache lines

Array in memory loading to cache in 2 operations

0 bytes

64 bytes

128 bytes
Memory alignment

Array in memory alignment to 64 bytes

64 byte cache lines

L2 CACHE

AVX

Array in L2 CACHE loading to AVX register in 2 operations
**SIMD datatypes**

- Intel® AVX 512:
  - Vector: 512 bit
  - Data types: 32 and 64 bit integers, 32 and 64 bit floats
  - VL: 8, 16
**SIMD vectorization**

- Usage of AVX vector operations for data processing can be achieved by:
  - Manual
  - Automatically by compiler
    - **Minimize dependencies among iteration**

```c
for (int i = 0; i < n; i++)
    float dy = p.y[j] - p.y[i];
```

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
</tbody>
</table>
IVDEP vs SIMD pragmas

Difference between IVDEP & SIMD pragmas

#pragma ivdep
  • Ignore vector dependencies (IVDEP):
    • Compiler ignore assumed but not proven dependencies for loop
  • Example:
    ```
    void foo(int *a, int k, int c, int m) {
      #pragma ivdep
      for (int i = 0; i < m; i++)
        a[i] = a[i + k] * c;
    }
    ```

#pragma simd
  • Aggressive version of IVEDP: Ignore all dependencies inside a loop
  • It’s an imperative that forces the compiler try everything to vectorize
  • Efficiency heuristic is ignored
  • It can vectorize code legally in some cases that wouldn’t be possible:
## SIMD vectorization: #pragma simd

### Example: without #pragma simd

```c
void add_floats(float *a, float *b, float *c, float *d, float *e, int n) {
    int i; for (i=0; i<n; i++) {
        a[i] = a[i] + b[i] + c[i] + d[i] + e[i];
    }
}
```

```
icc example1.c -nologo -Qvec-report2 example1.c
~\example1.c(3): (col. 3) remark: loop was not vectorized: existence of vector dependence.
```

### Example: with #pragma simd

```c
void add_floats(float *a, float *b, float *c, float *d, float *e, int n) {
    int i;
    #pragma simd
    for (i=0; i<n; i++) {
        a[i] = a[i] + b[i] + c[i] + d[i] + e[i];
    }
}
```

```
icc example1.c -nologo -Qvec-report2 example1.c
~\example1.c(4): (col. 3) remark: SIMD LOOP WAS VECTORIZED.
```
void bodyForce(BodySystem p, float dt, int n, int tileSize) {
    for (int tile = 0; tile < n; tile += tileSize) {
        int to = tile + tileSize;
        if (to > n) to = n;

        #pragma omp parallel for schedule(dynamic)
        for (int i = 0; i < n; i++) {
            float Fx = 0.0f; float Fy = 0.0f; float Fz = 0.0f;

            #pragma vector aligned
            #pragma simd
            for (int j = tile; j < to; j++) {
                float dy = p.y[j] - p.y[i];
                float dz = p.z[j] - p.z[i];
                float dx = p.x[j] - p.x[i];
                float distSqr = dx*dx + dy*dy + dz*dz + SOFTENING;
                float invDist = 1.0f / sqrtf(distSqr);
                float invDist3 = invDist * invDist * invDist;

                Fx += dx * invDist3; Fy += dy * invDist3; Fz += dz * invDist3;
            }

            p.vx[i] += dt*Fx; p.vy[i] += dt*Fy; p.vz[i] += dt*Fz;
        }
    }
}
Memory alignment and SIMD vectorization: compile & run

• Compile
  – #./build-nbody-align.sh

• Run
  – #./run-nbody-align.sh
Memory alignment and SIMD vectorization: performance results

![Bar chart showing interaction per second for different number of bodies with naive, SoA, Cache, and Aligned methods.](Image)