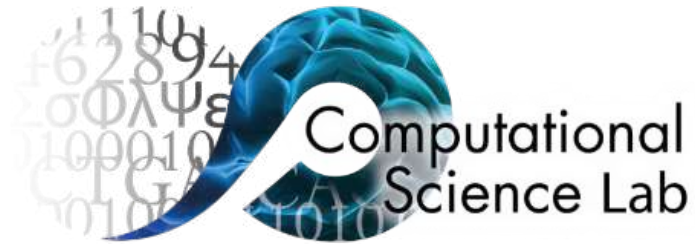


Performance optimizations in large-scale biomedical computations



Gábor Závodszy^{1,2}



¹University of Amsterdam, Computational Science Institute,
Amsterdam, The Netherlands

²Budapest University of Technology and Economics, Department of
Hydrodynamic Systems, Budapest, Hungary

Disclaimer

- **Performance** can refer to several things – in this talk I simply mean **wall clock run-time performance**

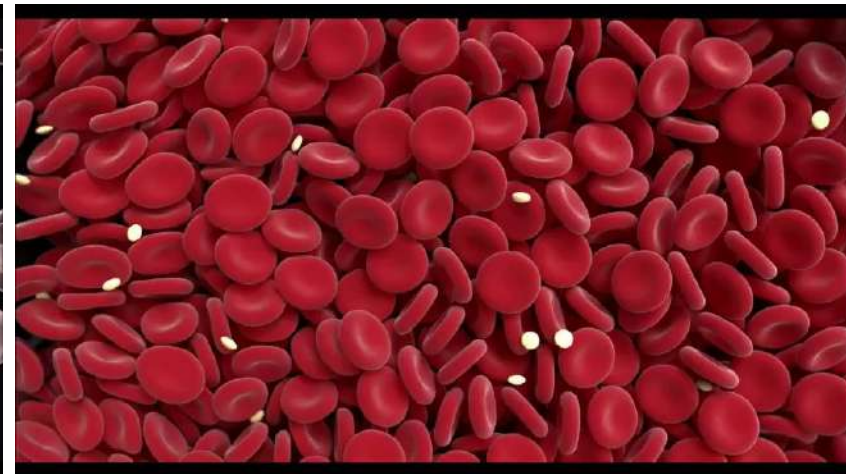
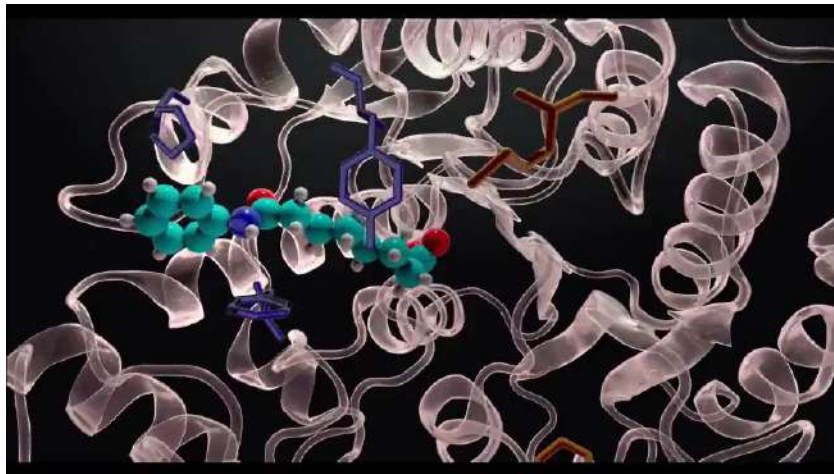
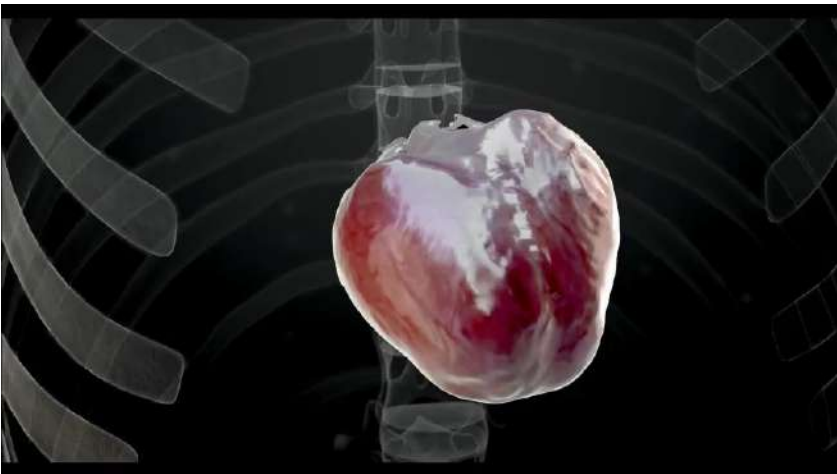
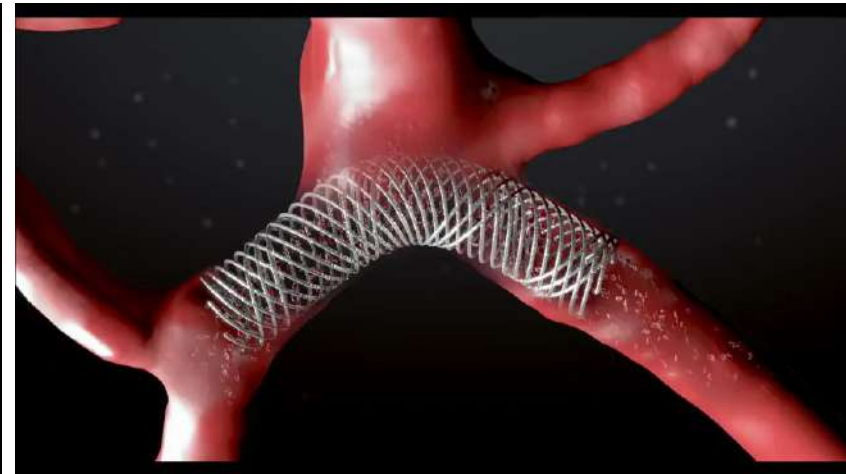
Outline

- Brief introduction to numerical models. What happens at large-scale deployments?
- Cherry-picking from the most often used techniques to improve computational performance.
 - These are just a few examples.
 - High-level overview.
 - From the perspective of a modeler.
- These techniques in the current talk are divided into three levels:
 - Low-level / architecture dependent programming techniques
 - Medium-level / algorithm implementation level techniques
 - High-level / conceptual-level considerations



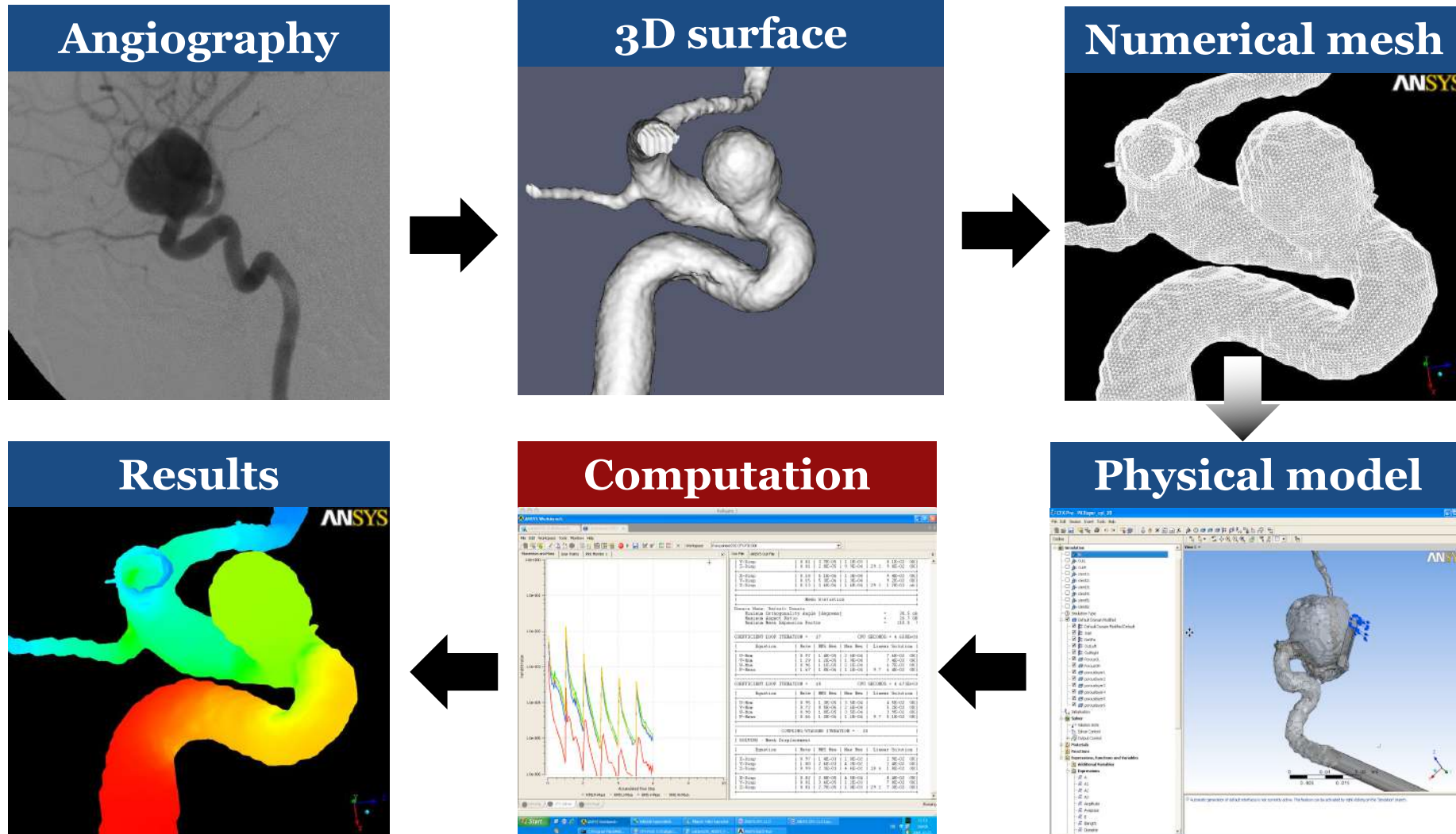
<https://youtu.be/1FvRSJ9W734>

Some example models used in research



Intro – numerical models

Modelling workflow



Numerical methods



Physical system

$$\left\{ \begin{array}{l} \rho \frac{\partial \mathbf{u}}{\partial t} + \rho(\mathbf{u} \cdot \nabla) \mathbf{u} - \nabla \cdot \boldsymbol{\sigma}(\mathbf{u}, p) = \mathbf{f} \\ \nabla \cdot \mathbf{u} = 0 \\ \mathbf{u} = \mathbf{g} \\ \boldsymbol{\sigma}(\mathbf{u}, p) \hat{\mathbf{n}} = \mathbf{h} \\ \mathbf{u}(0) = \mathbf{u}_0 \end{array} \right. \quad \text{PDE}$$

Analytic technique / approximation

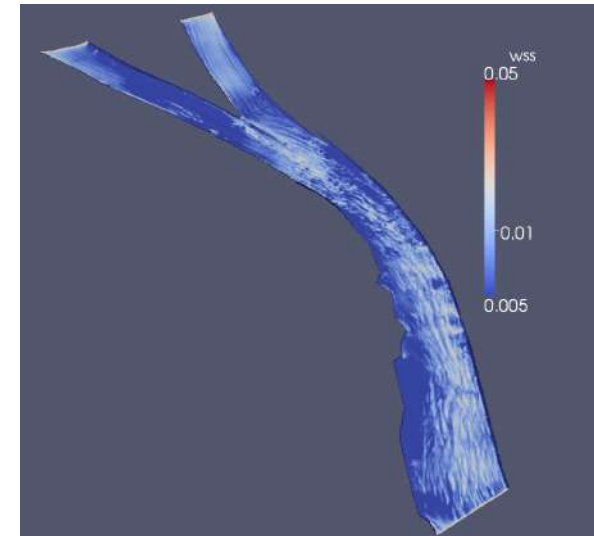
$$\text{ODE} \quad \begin{pmatrix} y_1^{(n)} \\ y_2^{(n)} \\ \vdots \\ y_m^{(n)} \end{pmatrix} = \begin{pmatrix} f_1(x, \mathbf{y}, \mathbf{y}', \mathbf{y}'', \dots, \mathbf{y}^{(n-1)}) \\ f_2(x, \mathbf{y}, \mathbf{y}', \mathbf{y}'', \dots, \mathbf{y}^{(n-1)}) \\ \vdots \\ f_m(x, \mathbf{y}, \mathbf{y}', \mathbf{y}'', \dots, \mathbf{y}^{(n-1)}) \end{pmatrix}$$

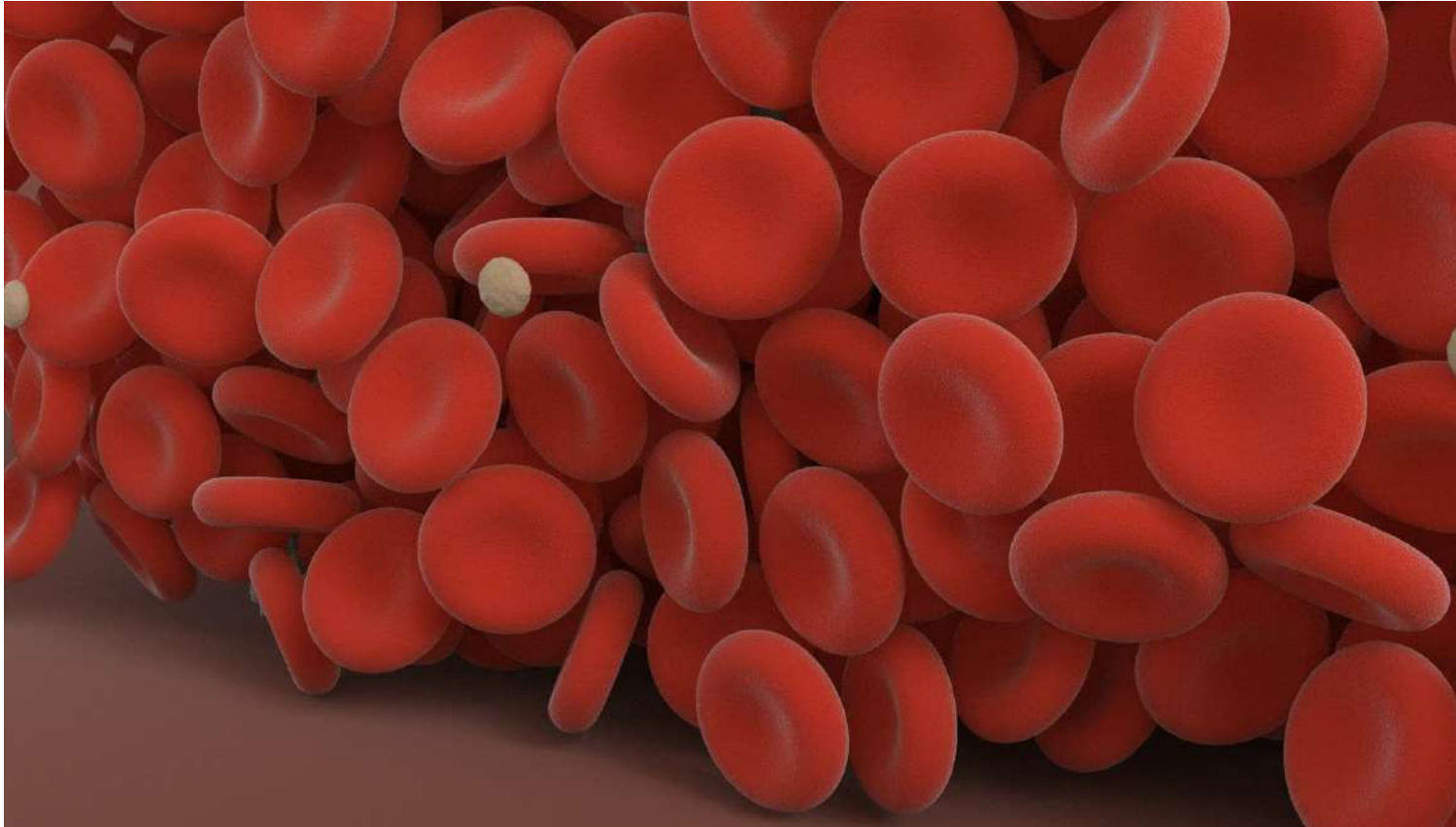
Numerical method

$$\text{Algebraic} \quad \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

where a_{ij} and b_i are constants, $i = 1, 2, \dots, m, j = 1, 2, \dots, n$.

Numerical solution



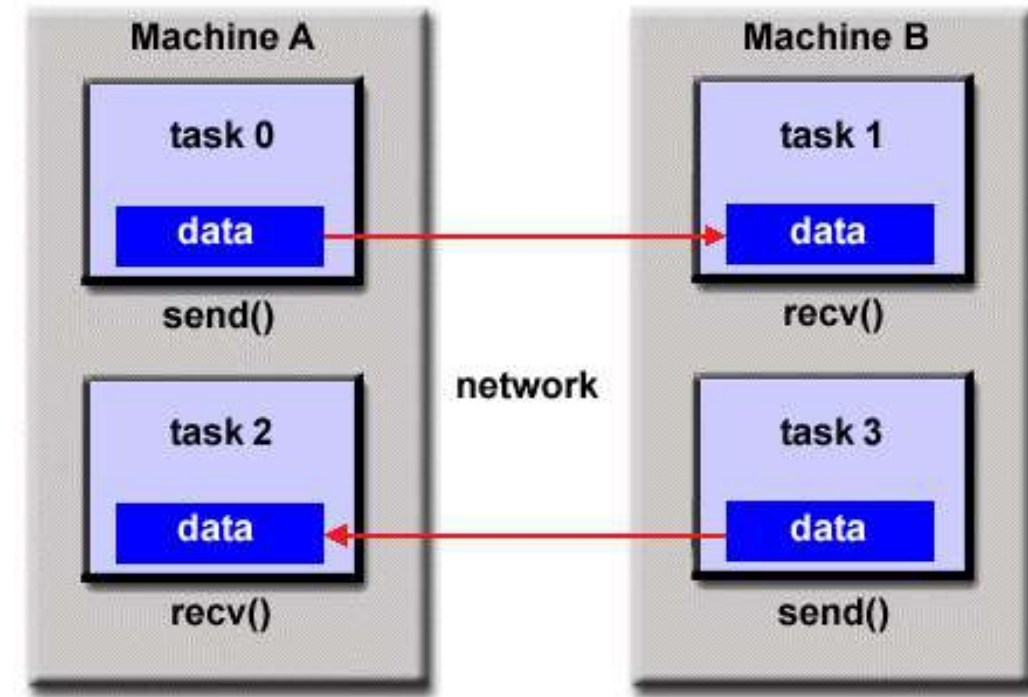


- Open source (www.hemocell.eu)
- Fully validated
- Currently used for research in:
 - Malaria
 - Diabetes
 - Hemolysis
 - Retinal aneurysms
 - Thrombus formation
 - Cellular and drug transport
 - Platelet aggregation

This model demonstrates the following characteristics:

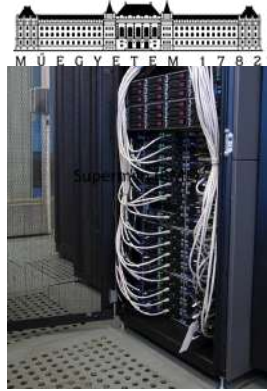
- A set of tasks that use their own local memory during computation. Multiple tasks can reside on the same physical machine and/or across an arbitrary number of machines.
- Tasks exchange data through communications by sending and receiving messages.
- Data transfer usually requires cooperative operations to be performed by each process. For example, a send operation must have a matching receive operation.

Note: in some cases shared memory parallelism can be preferable. I.e.: OpenMP can be more portable (GPU support), faster to implement (decorator over a loop).



Barney, B. (2010). Introduction to parallel computing. Lawrence Livermore National Laboratory, 6(13), 10.

HPC deployment



Superman (BME, Budapest)



Lomonosov (MSU, Moscow)



Marenostrum (BSC, Barcelona)



SGI Altrix (QUT, Brisbane)



Eagle (PSNC, Poznan)



Cartesius and Lisa (SURFsara, Amsterdam)



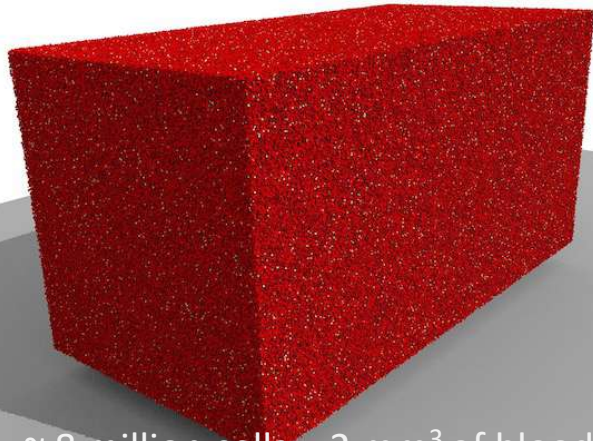
Sanam (KACST)



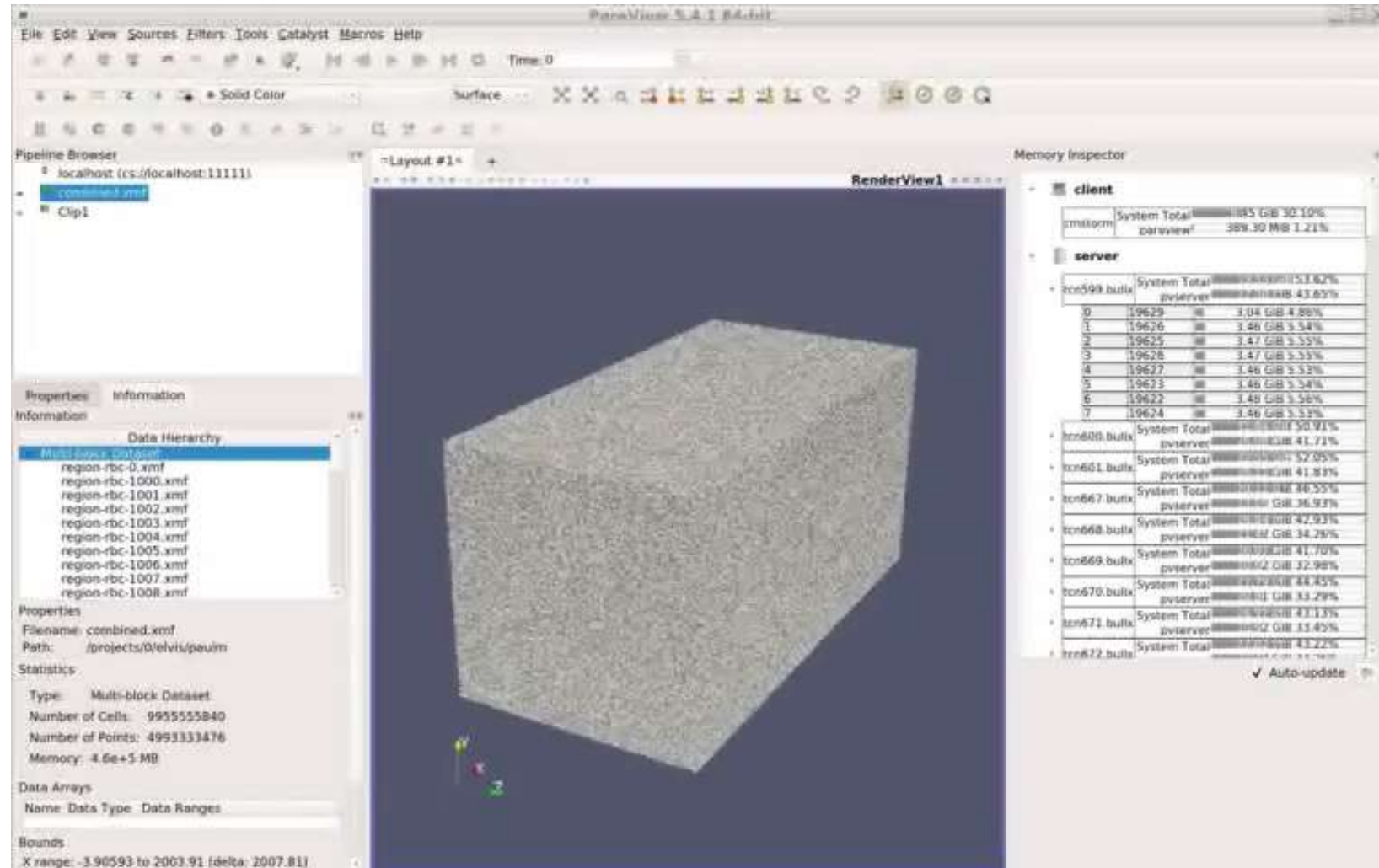
Supermuc (LRZ, Munich)



Aspire I (NSCC, Singapore)



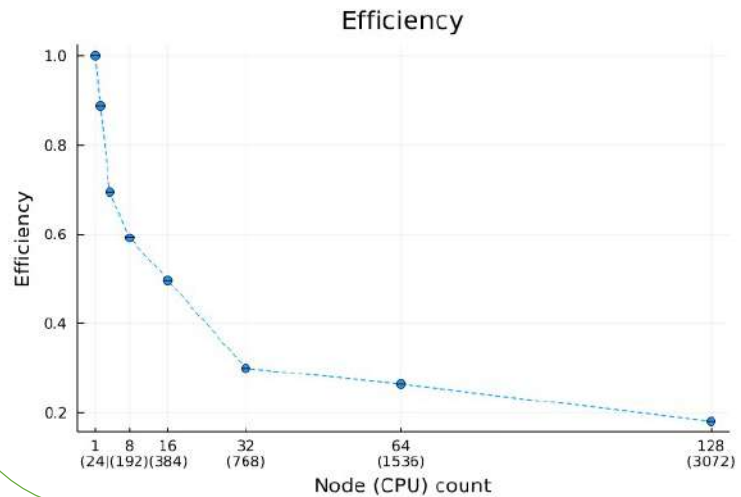
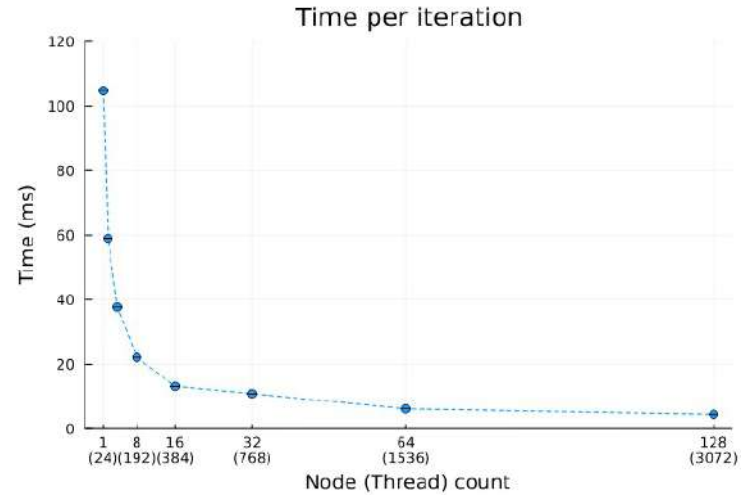
~ 8 million cells = 2 mm³ of blood



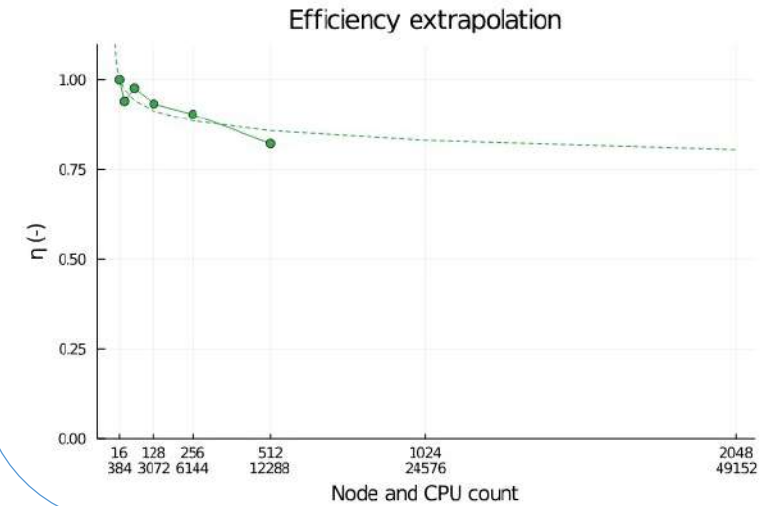
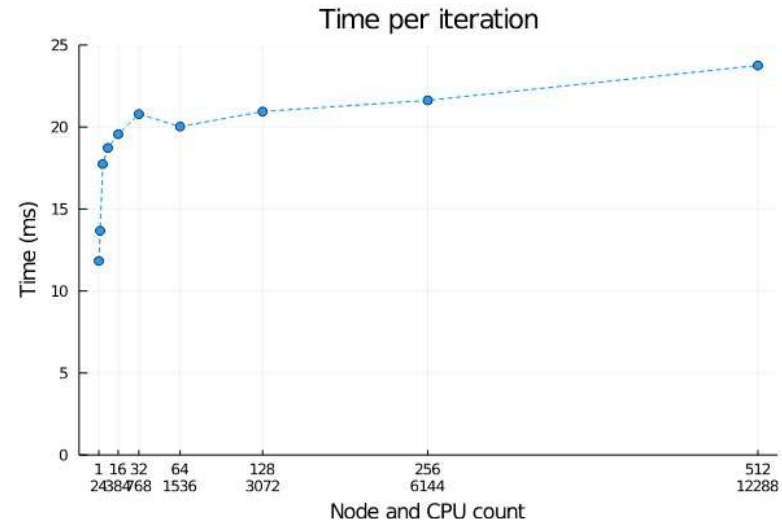
Paul Melis, SURF

HPC – Scaling and performance

Strong-scaling



Weak-scaling



Why is it important to improve scaling?
Which scaling is more important?
Which on is more difficult to improve?

Low-level techniques

(Implementation techniques targeting architectural properties = close to metal)

Algebraic operations: $x / 2$ vs. $0.5 * x$?

		Intel P4 F4		Intel Core 2		Intel NHM		Intel SBR		Intel HWL		Intel BWL		Intel SKL		AMD K8-K9		AMD K10		AMD BD2		AMD BD4		AMD Zn1		AMD Zn2	
		L64	T64	L64	T64	L64	T64	L64	T64	L64	T64	L64	T64	L64	T64	L64	T64	L64	T64	L64	T64	L64	T64	L64	T64	L64	T64
add	r,ri	1	2.5	1	3	1	3	1	3	1	4	1	4	1	4	1	3	1	3	1	2	1	3.5	1	4	1	4
sub	r,ri	1	2.5	1	3	1	3	1	3	1	4	1	4	1	4	1	3	1	3	1	2	1	3.5	1	4	1	4
and	r,r	1	2	1	3	1	3	1	3	1	4	1	4	1	4	1	3	1	3	1	2	1	3.5	1	4	1	4
or	r,r	1	2	1	3	1	3	1	3	1	4	1	4	1	4	1	3	1	3	1	2	1	3.5	1	4	1	4
xor	r,r	1	2	1	3	1	3	1	3	1	4	1	4	1	4	1	3	1	3	1	2	1	3.5	1	4	1	4
inc	r	1	1	1	3	1	3	1	3	1	4	1	4	1	4	1	3	1	3	1	2	1	3.5	1	4	1	4
dec	r	1	1	1	3	1	3	1	3	1	4	1	4	1	4	1	3	1	3	1	2	1	3.5	1	4	1	4
neg	r	1	2	1	3	1	3	1	3	1	4	1	4	1	4	1	3	1	3	1	2	1	3.5	1	4	1	4
not	r	1	1.7	1	3	1	3	1	3	1	4	1	4	1	4	1	3	1	3	1	2	1	3.5	1	4	1	4
imul	r,ri	10	1/2	5	1/2	3	1	3	1	3	1	3	1	3	1	4	1/2	4	1/2	6	1/4	6	1/4	3	1	3	1
mul	r	12	1/10	8	1/4	10 ⁵	1/2	4 ⁶	1	3	1	3	1	3	1	5 ¹	1/2	5 ¹	1/2	7 ^x	1/4	7 ^x	1/4	4 ⁶	1/2	4 ⁶	1
mulx	r,r,r	-	-	-	-	-	-	3	1	3	1	3	1	3	1	-	-	-	-	-	-	-	-	4 ⁶	1/2	4 ⁶	1
div	r	161 ²	1/151	116 ⁴		89		92		95		94		86		71	1/71	77	1/77	76	1/76	76	1/76	46	1/46	46	1/46
adc	r,ri	10	1/3	2	1	2	1	2 ⁷	1	2 ⁷	1	1	1	1	1	1	3	1	3	1	2	1		1		1	
sbb	r,ri	10	1/3	2	1	2	1	2 ⁷	1	2 ⁷	1	1	1	1	1	1	3	1	3	1	2	1		1		1	

Note: ***0.5** has a different round-off error than **/ 2**!

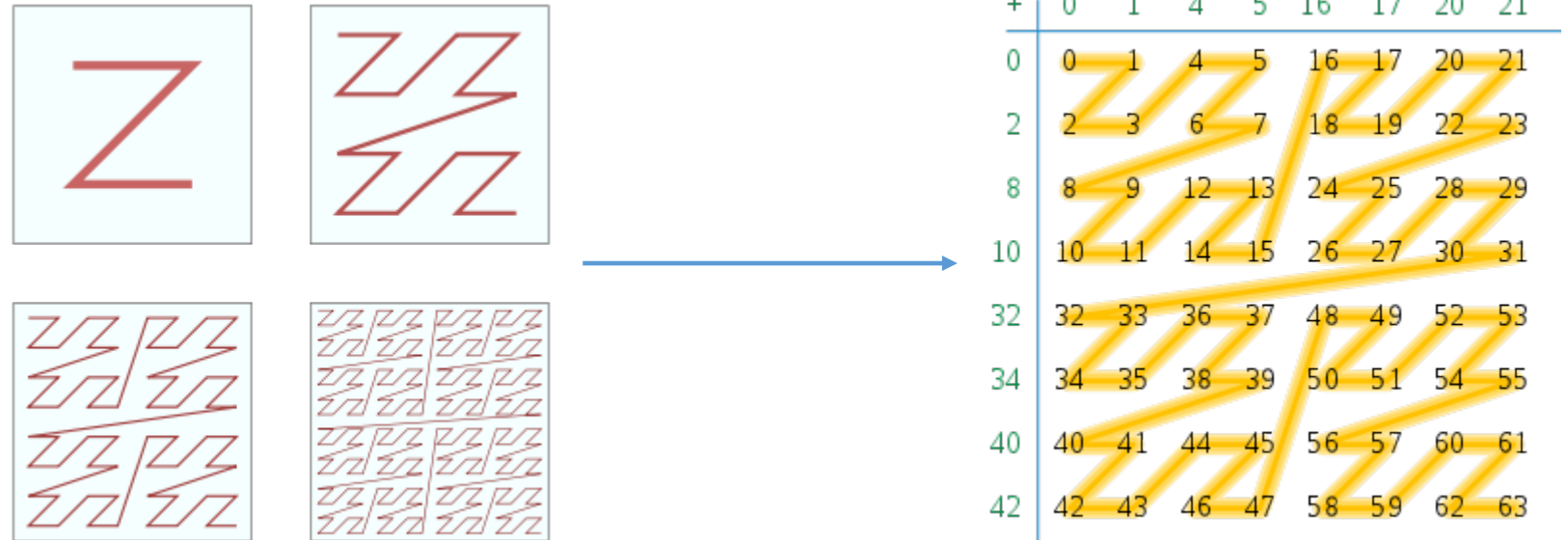
<https://gmplib.org/~tege/x86-timing.pdf>

Improving data locality – spatial indexing

There are numerous other space-filling curves.

Heavily used e.g. in GPU related applications.

It can improve cache hits significantly.



```
struct {  
    uint8_t r, g, b;  
} AoS[N];
```

```
struct {  
    uint8_t r[N];  
    uint8_t g[N];  
    uint8_t b[N];  
} SoA;
```

- Appropriate data structures (e.g. AoS vs. SoA) that improve data locality.
This depends on the access pattern of the values.
- Loop optimizations (e.g. via loop transformations).
Via various transformations fusing loops together can improve performance substantially, plus it allows for additional optimizations (e.g., data reuse).
- Aim for data reuse.
Commonly used values can be stored instead of being recomputed.

Medium-level techniques

(improvements in the implementation methods)

Metaprogramming example

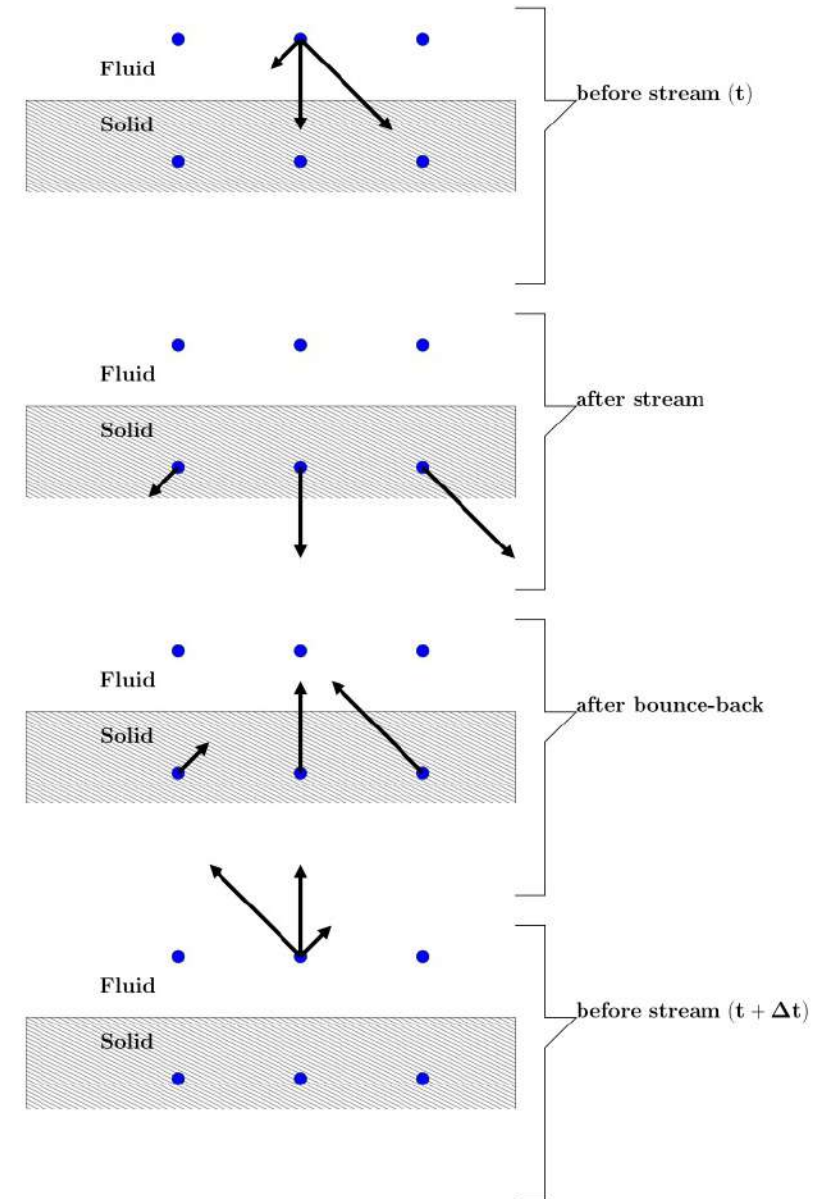
Can be used to (amongst others):

- Keep the model description close to math form.
- Generate code from symbolic math expressions conveniently.
- Generate code for multiple backends (e.g. CPU, GPU).
- Keep the code flexible and general.

Two example algorithms (from a fluid flow simulation):

A.) streaming data on multiple discretized grids

B.) implement a mathematical vector equation



Metaprogramming allows for general code that works in 2D or 3D.

Mako code:

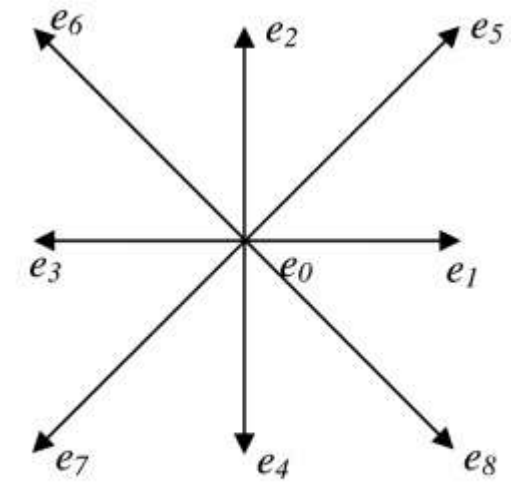
```
${device_func} inline void bounce_back(Dist *fi)
{
    float t;

    %for i in sym.bb_swap_pairs(grid):
        t = fi->${grid.idx_name[i]};
        fi->${grid.idx_name[i]} = fi->${grid.idx_name[grid.idx_opposite[i]]};
        fi->${grid.idx_name[grid.idx_opposite[i]]} = t;
    %endfor
}
```

Metaprogramming implementation of the bounce-back method in Python, using the Mako templating language. This technique is trivial in some languages (e.g., lisp), and quite difficult in others (e.g., fortran).

CUDA C code, D2Q9 grid:

```
__device__ inline void bounce_back(Dist * fi)
{
    float t;
    t = fi->fE;
    fi->fE = fi->fW;
    fi->fW = t;
    t = fi->fN;
    fi->fN = fi->fS;
    fi->fS = t;
    t = fi->fNE;
    fi->fNE = fi->fSW;
    fi->fSW = t;
    t = fi->fNW;
    fi->fNW = fi->fSE;
    fi->fSE = t;
}
```



D2Q9

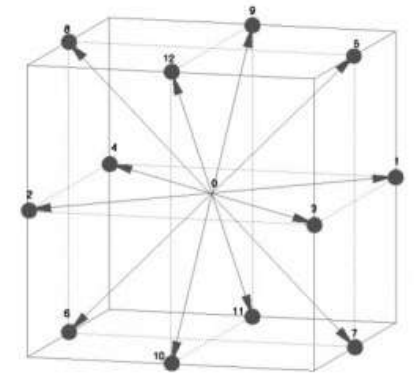
CUDA C code, D3Q13 grid:

```
__device__ inline void bounce_back(Dist * fi)
{
    float t;
    t = fi->fNE;
    fi->fNE = fi->fSW;
    fi->fSW = t;
    t = fi->fSE;
    fi->fSE = fi->fNW;
    fi->fNW = t;
    t = fi->fTE;
    fi->fTE = fi->fBW;
    fi->fBW = t;
    t = fi->fBE;
    fi->fBE = fi->fTW;
    fi->fTW = t;
}
```

...

```
t = fi->fTN;
fi->fTN = fi->fBS;
fi->fBS = t;
t = fi->fBN;
fi->fBN = fi->fTS;
fi->fTS = t;
```

}



D3Q13

Metaprogramming allows for abstractions, to keep code close to math equations

$$f_i^{eq}(\vec{x}, t) = w_i \rho [1 + 3(\vec{e}_i \cdot \vec{u}) + \frac{9}{2}(\vec{e}_i \cdot \vec{u})^2 - \frac{3}{2}\vec{u}^2]$$

```
def bgk_equilibrium(grid, rho=None):
    out = []

    if rho is None:
        rho = S.rho

    for i, ei in enumerate(grid.basis):
        t = (grid.weights[i] * rho * (1 +
            3*ei.dot(grid.v) +
            Rational(9, 2) * (ei.dot(grid.v))**2 -
            Rational(3, 2) * grid.v.dot(grid.v)))

        out.append(t)

    return out
```

Code close to math equations

$$f_i^{eq}(\vec{x}, t) = w_i \rho [1 + 3(\vec{e}_i \cdot \vec{u}) + \frac{9}{2}(\vec{e}_i \cdot \vec{u})^2 - \frac{3}{2}\vec{u}^2]$$

```
def bgk_equilibrium(grid, rho=None):  
    out = []  
  
    if rho is None:  
        rho = S.rho  
  
    for i, ei in enumerate(grid.basis):  
        t = (grid.weights[i] * rho * (1 +  
            3*ei.dot(grid.v) +  
            Rational(9, 2) * (ei.dot(grid.v))**2 -  
            Rational(3, 2) * grid.v.dot(grid.v)))  
  
        out.append(t)  
  
    return out
```

$$f_i^{eq}(\vec{x}, t) = w_i \rho [1 + 3(\vec{e}_i \cdot \vec{u}) + \frac{9}{2}(\vec{e}_i \cdot \vec{u})^2 - \frac{3}{2}\vec{u}^2]$$

```
def bgk_equilibrium(grid, rho=None):
    out = []

    if rho is None:
        rho = S.rho

    for i, ei in enumerate(grid.basis):
        t = (grid.weights[i] * rho * (1 +
            3*ei.dot(grid.v) +
            Rational(9, 2) * (ei.dot(grid.v))**2 -
            Rational(3, 2) * grid.v.dot(grid.v)))

        out.append(t)

    return out
```


Code close to math equations

$$f_i^{eq}(\vec{x}, t) = w_i \rho [1 + 3(\vec{e}_i \cdot \vec{u}) + \frac{9}{2}(\vec{e}_i \cdot \vec{u})^2 - \frac{3}{2}\vec{u}^2]$$

```
def bgk_equilibrium(grid, rho=None):
    out = []

    if rho is None:
        rho = S.rho

    for i, ei in enumerate(grid.basis):
        t = (grid.weights[i] * rho * (1 +
            3*ei.dot(grid.v) +
            Rational(9, 2) * (ei.dot(grid.v))**2 -
            Rational(3, 2) * grid.v.dot(grid.v)))

        out.append(t)

    return out
```

$$f_i^{eq}(\vec{x}, t) = w_i \rho [1 + 3(\vec{e}_i \cdot \vec{u}) + \frac{9}{2}(\vec{e}_i \cdot \vec{u})^2 - \frac{3}{2}\vec{u}^2]$$

```
def bgk_equilibrium(grid, rho=None):
    out = []

    if rho is None:
        rho = S.rho

    for i, ei in enumerate(grid.basis):
        t = (grid.weights[i] * rho * (1 +
            3*ei.dot(grid.v) +
            Rational(9, 2) * (ei.dot(grid.v))**2 -
            Rational(3, 2) * grid.v.dot(grid.v)))

        out.append(t)

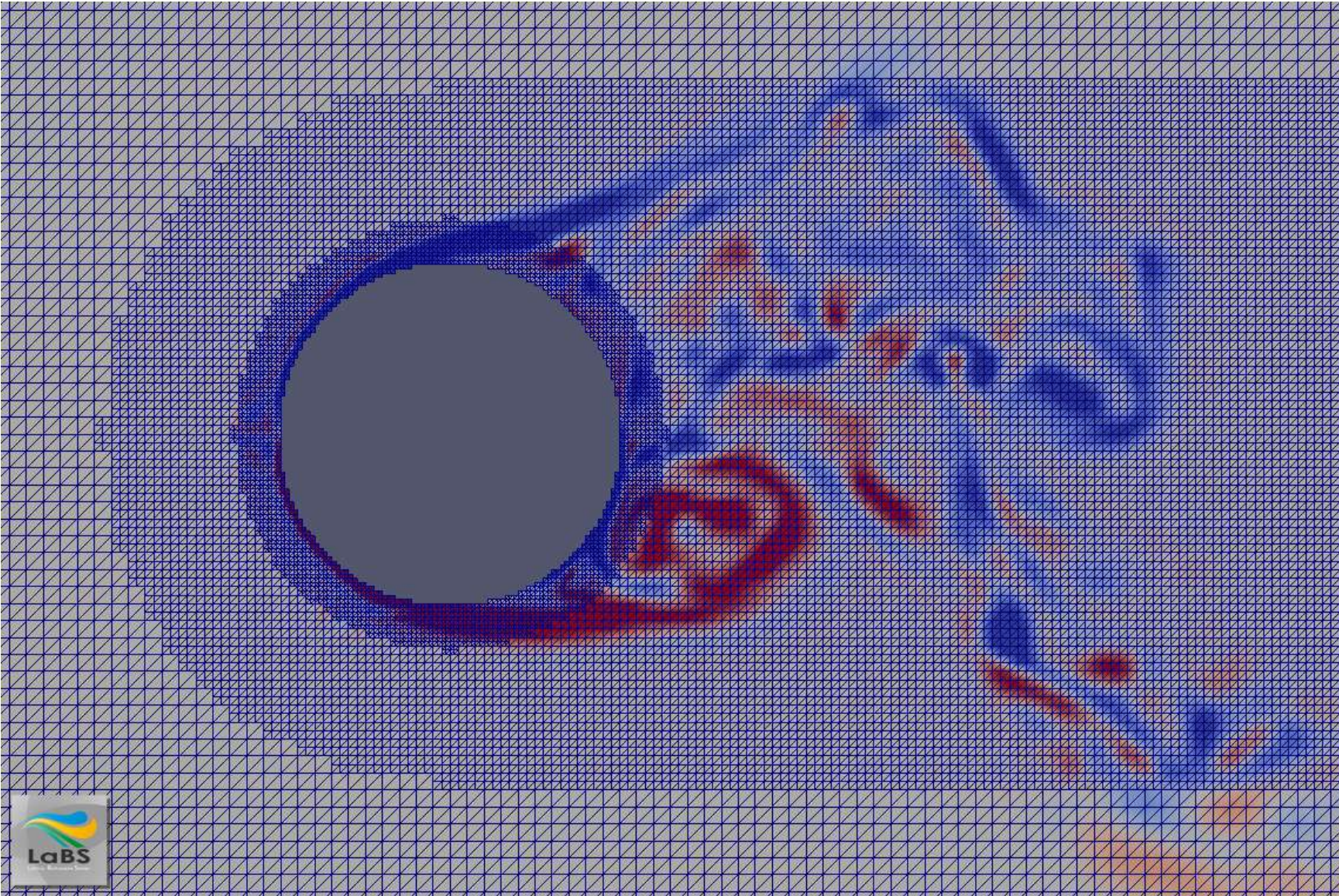
    return out
```

$$f_i^{eq}(\vec{x}, t) = w_i \rho [1 + 3(\vec{e}_i \cdot \vec{u}) + \frac{9}{2}(\vec{e}_i \cdot \vec{u})^2 - \frac{3}{2}\vec{u}^2]$$

The generated code:

```
freq0.fC = rho / 3 + rho * (-3 * v0[0] * v0[0] / 2 - 3 * v0[1] * v0[1] / 2 - 3 * v0[2] * v0[2] / 2) / 3;  
freq0.fE = rho / 18 + rho * (3 * v0[0] * (1 + v0[0]) - 3 * v0[1] * v0[1] / 2 - 3 * v0[2] * v0[2] / 2) / 18;  
freq0.fW = rho / 18 + rho * (-3 * v0[0] * (1 - v0[0]) - 3 * v0[1] * v0[1] / 2 - 3 * v0[2] * v0[2] / 2) / 18;  
freq0.fN = rho / 18 + rho * (3 * v0[1] * (1 + v0[1]) - 3 * v0[0] * v0[0] / 2 - 3 * v0[2] * v0[2] / 2) / 18;  
freq0.fS = rho / 18 + rho * (-3 * v0[1] * (1 - v0[1]) - 3 * v0[0] * v0[0] / 2 - 3 * v0[2] * v0[2] / 2) / 18;  
freq0.fT = rho / 18 + rho * (3 * v0[2] * (1 + v0[2]) - 3 * v0[0] * v0[0] / 2 - 3 * v0[1] * v0[1] / 2) / 18;  
freq0.fB = rho / 18 + rho * (-3 * v0[2] * (1 - v0[2]) - 3 * v0[0] * v0[0] / 2 - 3 * v0[1] * v0[1] / 2) / 18;  
freq0.fNE = rho / 36 + rho * (3 * v0[0] * (1 + v0[0]) + 3 * v0[1] * (1 + v0[1]) + 3 * v0[2] * v0[2] / 2) / 36;  
freq0.fNW = rho / 36 + rho * (-3 * v0[0] * (1 - v0[0]) + 3 * v0[1] * (1 + v0[1]) - 3 * v0[2] * v0[2] / 2) / 36;  
freq0.fSE = rho / 36 + rho * (-3 * v0[1] * (1 - v0[1]) + 3 * v0[0] * (1 + v0[0]) - 3 * v0[2] * v0[2] / 2) / 36;  
freq0.fSW = rho / 36 + rho * (-3 * v0[0] * (1 - v0[0]) - 3 * v0[1] * (1 - v0[1]) - 3 * v0[2] * v0[2] / 2) / 36;  
:  
:
```


Grid refinement - static

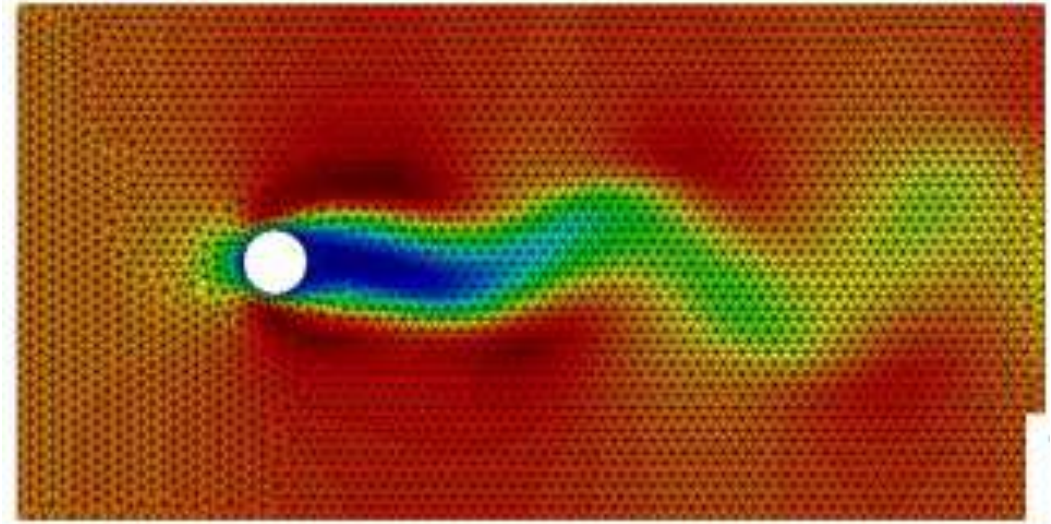
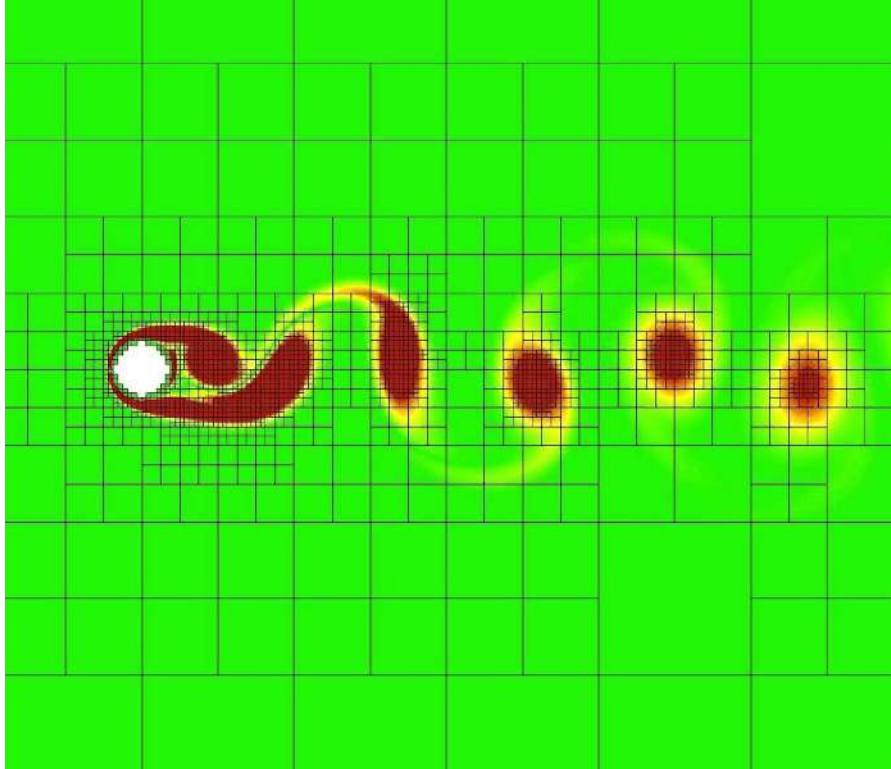


- We employ this method
- to increase accuracy or
 - to increase performance.

Think about what this means for parallel programming!

- The load-imbalance can be mitigated at the initial domain decomposition.
- At the cost of significant complexity!

Grid refinement - adaptive



- Dynamic (adaptive) grid refinement presents dynamic load-imbalance challenges!
- This is an order of magnitude more difficult problem, that is currently heavily researched.

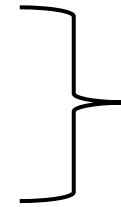
High-level techniques

(conceptual level of the numerical model – we use information about the specific problem being solved)

Dynamic iteration step-size in a coupled simulation

Time steps:

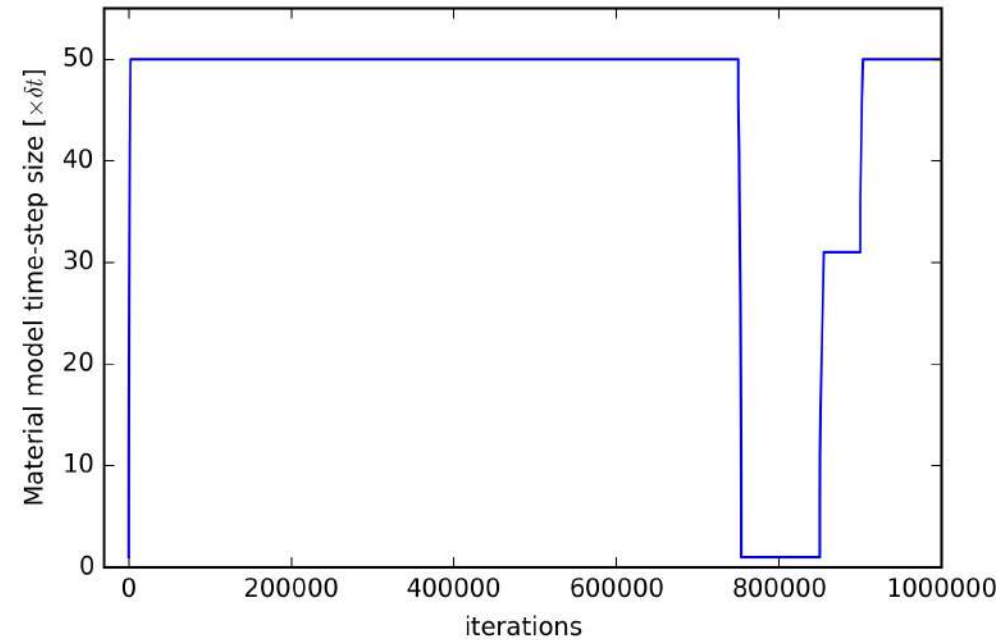
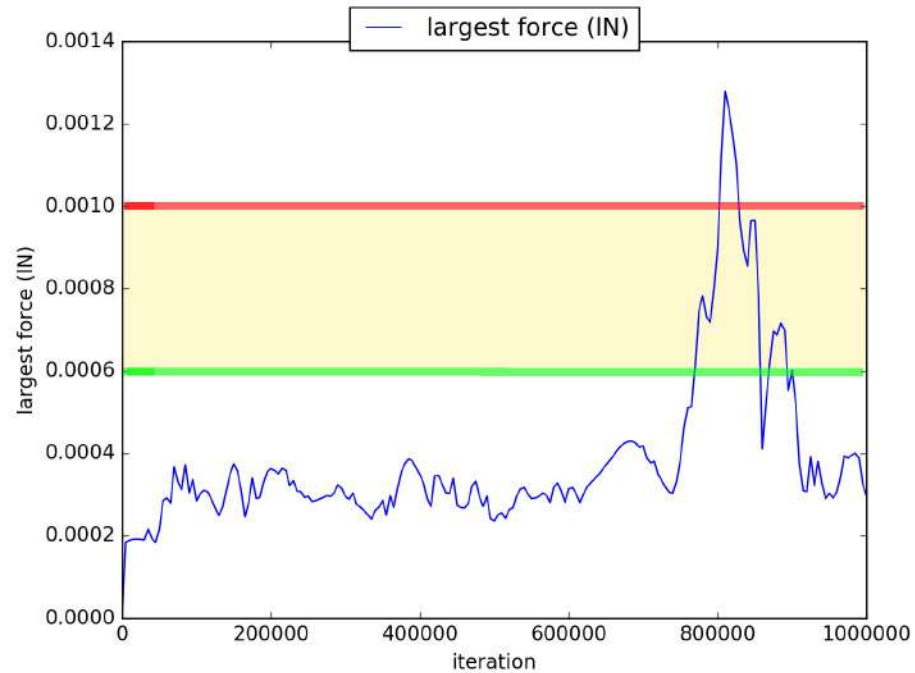
- Fluid field (LBM): Δt $\sim 1e-7$ s
- FSI coupling (IBM): Δt $\sim 1-6 \Delta t$
- Cell model (DEM): Δt $\sim 1-100 \Delta t$



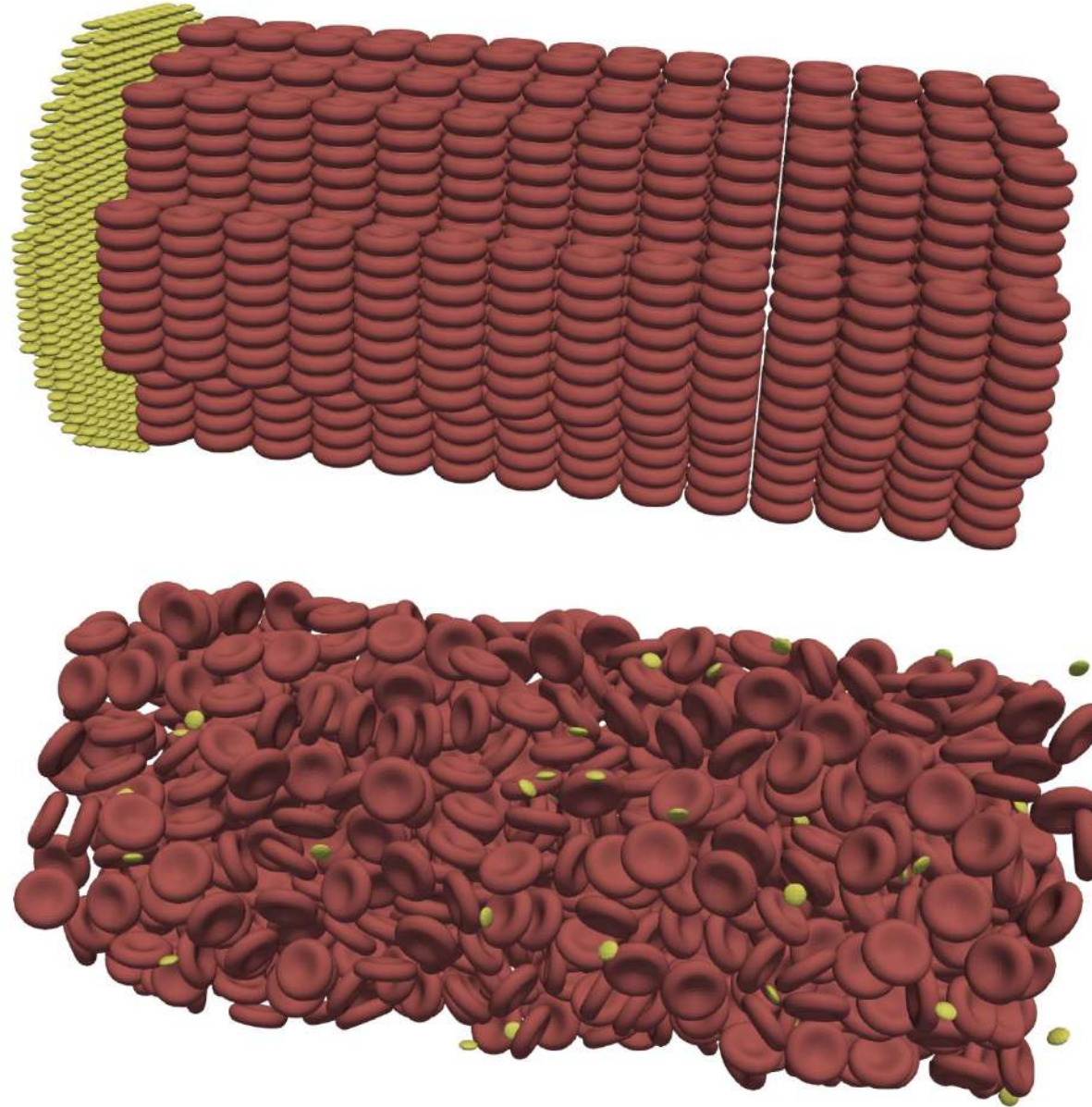
$$\Delta t \leq \Delta t \leq \Delta t$$

Technically it is an adaptive grid in time with different resolution for the different coupled components.

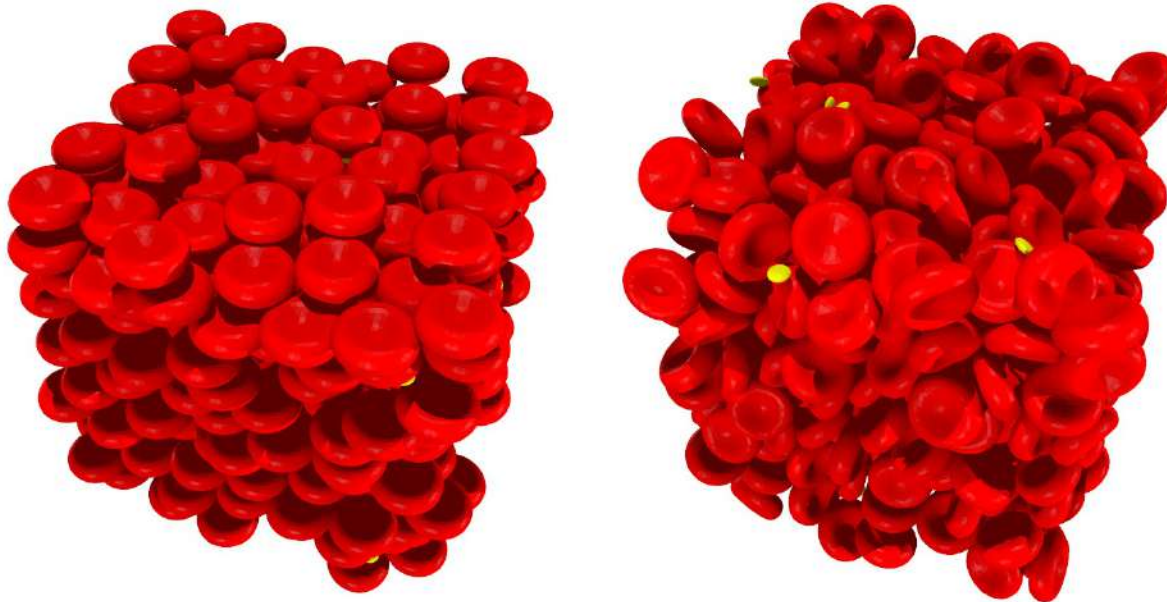
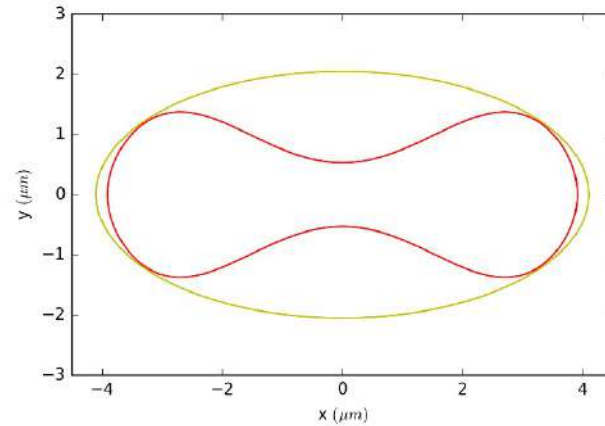
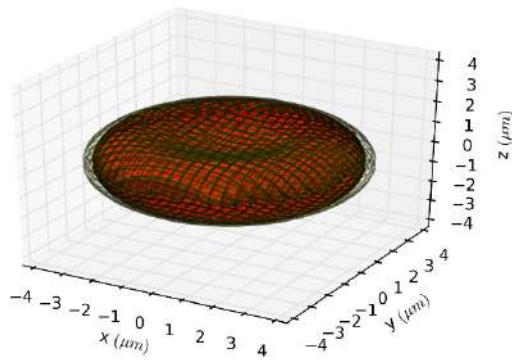
Δt :



Initial condition optimizations – better mixed cells



Fast way to compute the initial cell packing



Force-bias method:

$$\vec{F}_{ij} = \delta_{ij} p_{ij} \frac{\vec{r}_j - \vec{r}_i}{|\vec{r}_j - \vec{r}_i|}$$

- Potential function is proportional to overlapping volumes
- Regular grid space-partitioning
- OpenMP implementation
- On a 16 core machine it can position millions of cells at 40% hematocrit within an hour of computation wall-time

Gain through some optimization

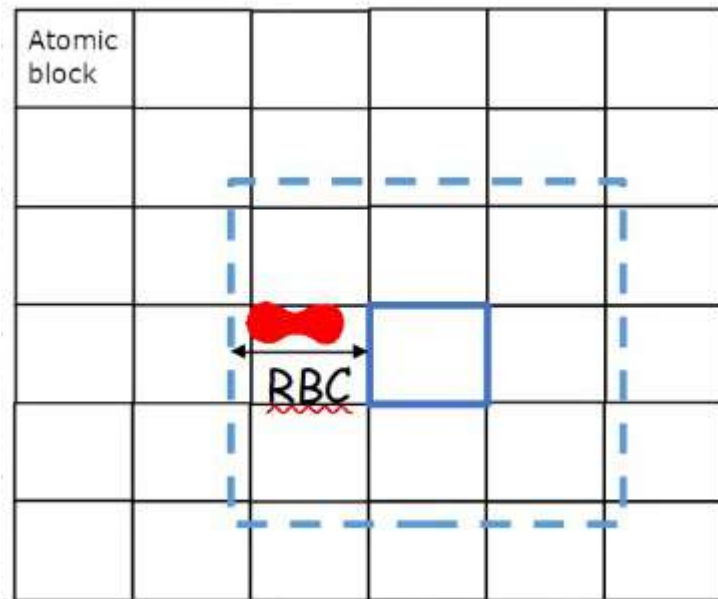
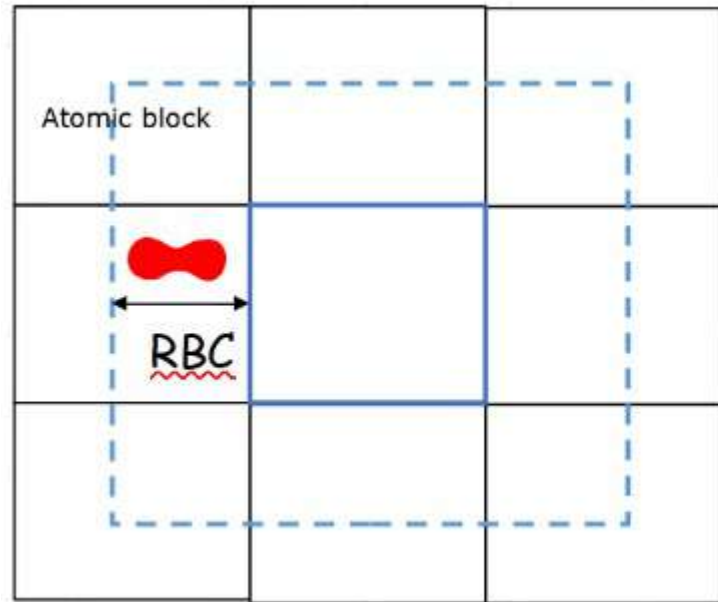
	Wall-time / iteration	Warmup iterations	Simulation iterations	Total time (expected)
original	2.3 s	10M	2M	319.5 days
+ low-level code opt.	0.9 s	10M	2M	125 days
+ better initial conditions	0.9 s	1M	2M	31.25 days
+ adaptive time steps	0.06 s	1M	2M	2.1 days

Conceptual level optimization (15x). **Think before you optimize!**

Domain decomposition of coupled codes



Communication as a bottleneck



The communication envelope size is constant (dictated by the simulated system).

The number of neighbours grow as $(2N+1)^3-1$ in a 3D grid, where N is the number of neighbours in one direction.

From $N=1$ to $N=2 \rightarrow 124-26 = 98$ new neighbours.

This is rather typical with simulations, and can cause a serious bottleneck for performance scaling. On an HPC this is a serious issue!

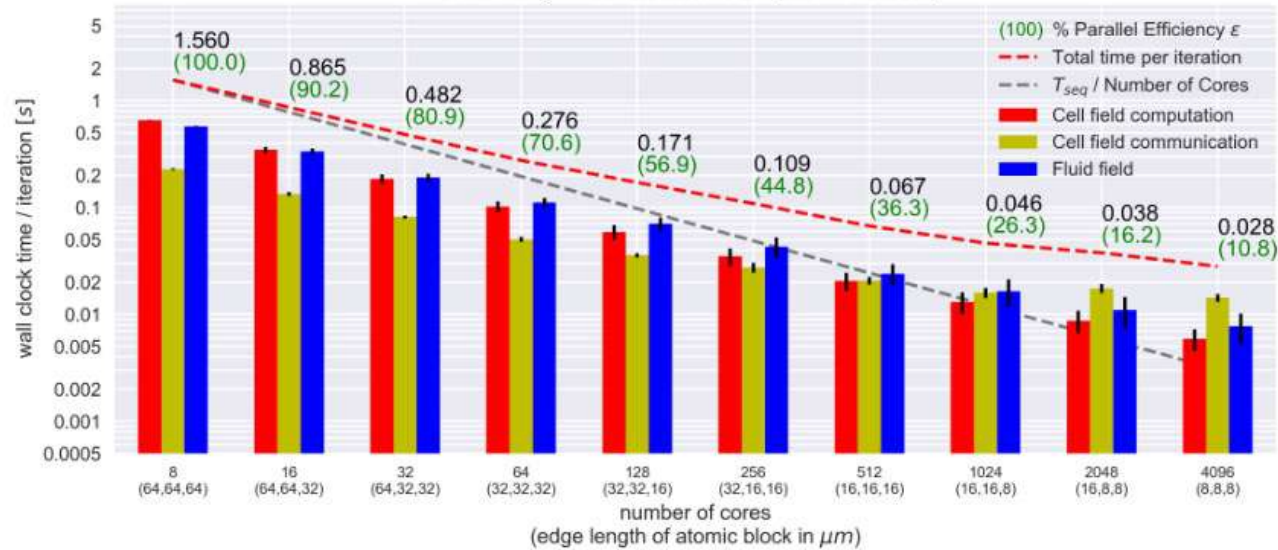
(Think! Is this a strong- or weak-scaling issue?)

In HemoCell to mitigate this particular issue we implemented a two-step communication:

1. Distribute what we specifically need from neighbours
2. Transfer only that

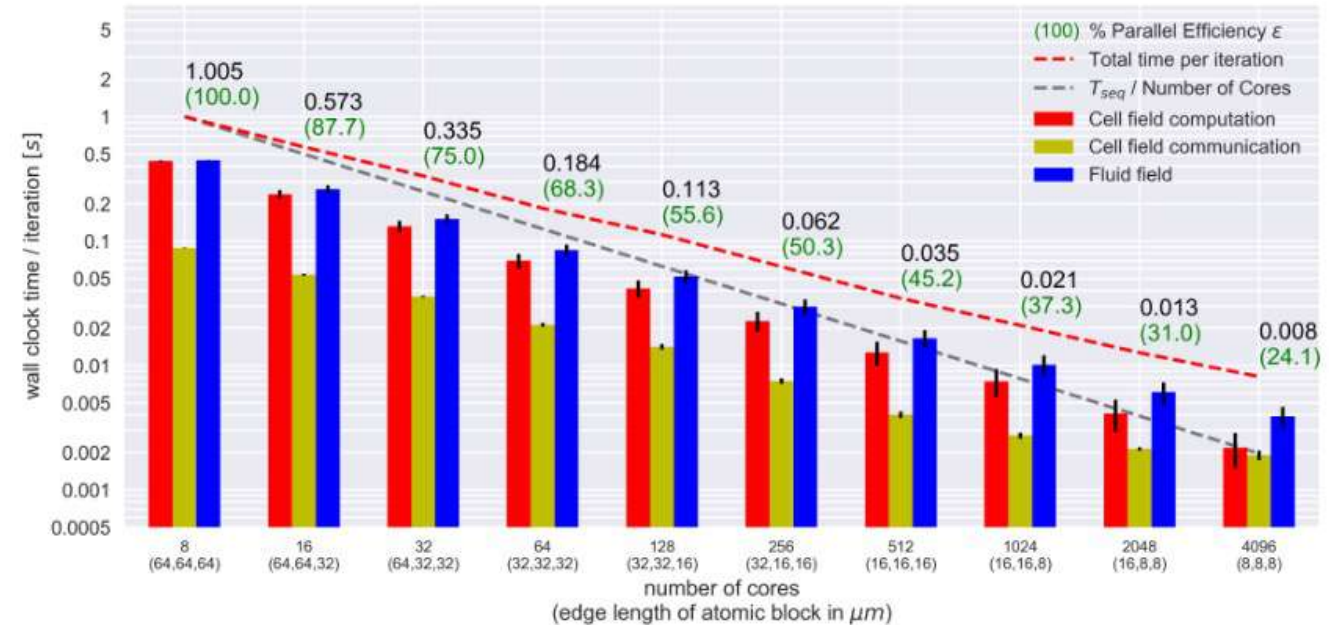
Communication as a bottleneck for scaling

Parallel efficiency of a simulation in a $2097152 \mu\text{m}^3$ domain containing 7736 red blood cells (33% hematocrit)

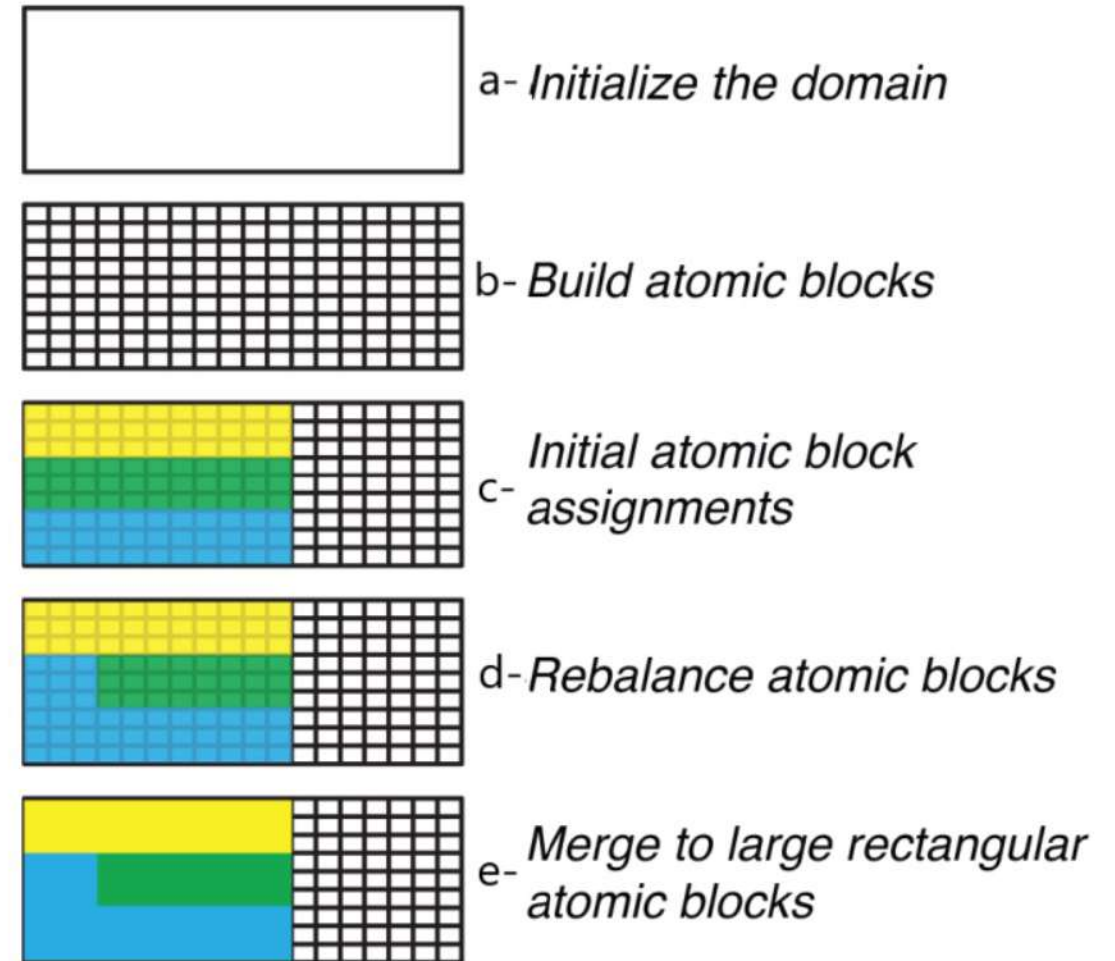
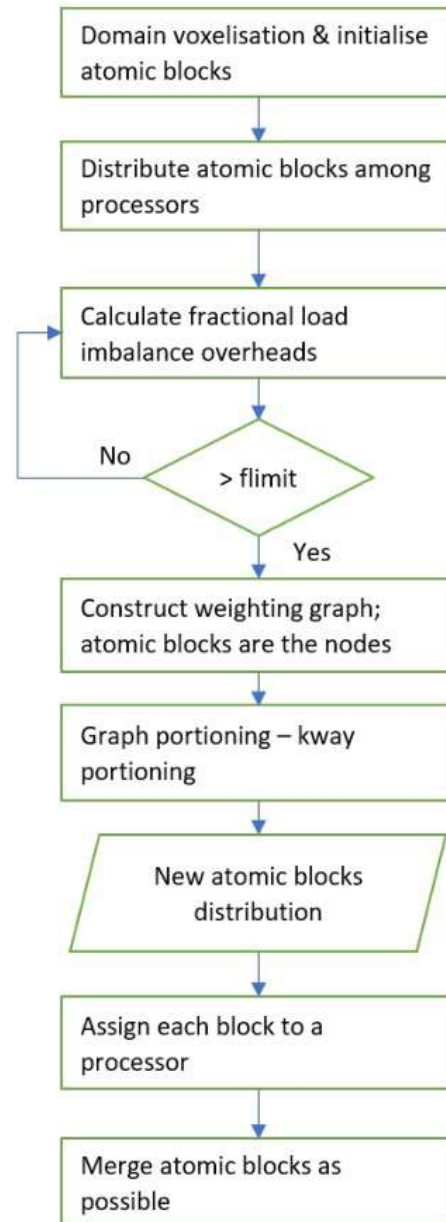


<- Single-step communication

Two-step communication ->

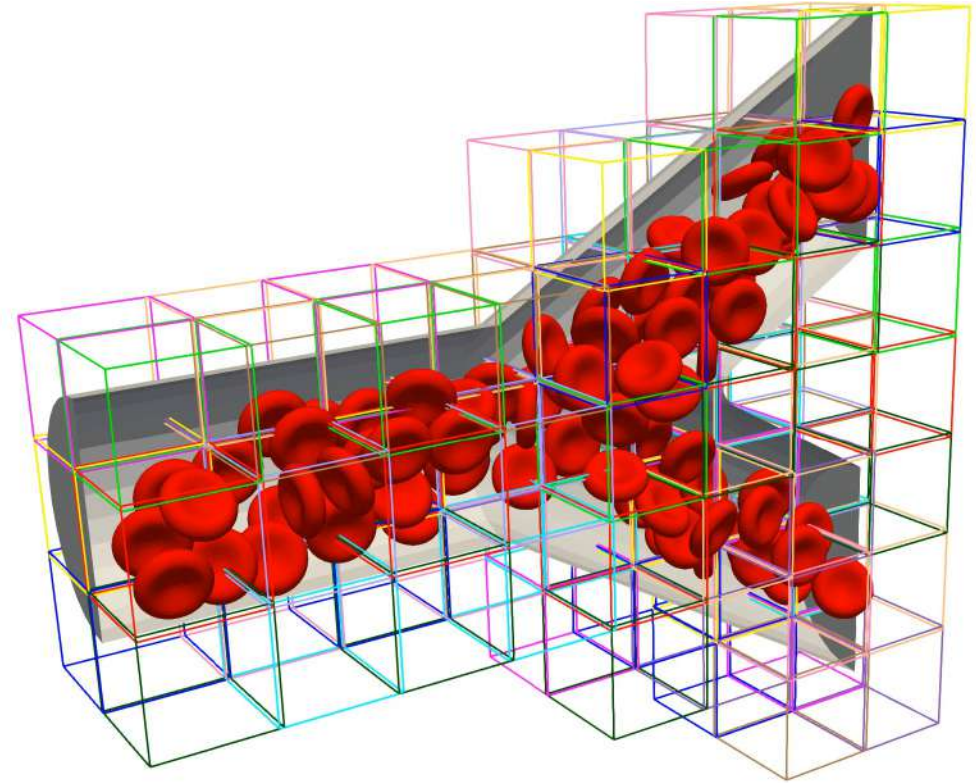


Load-balancing spatial domain distribution

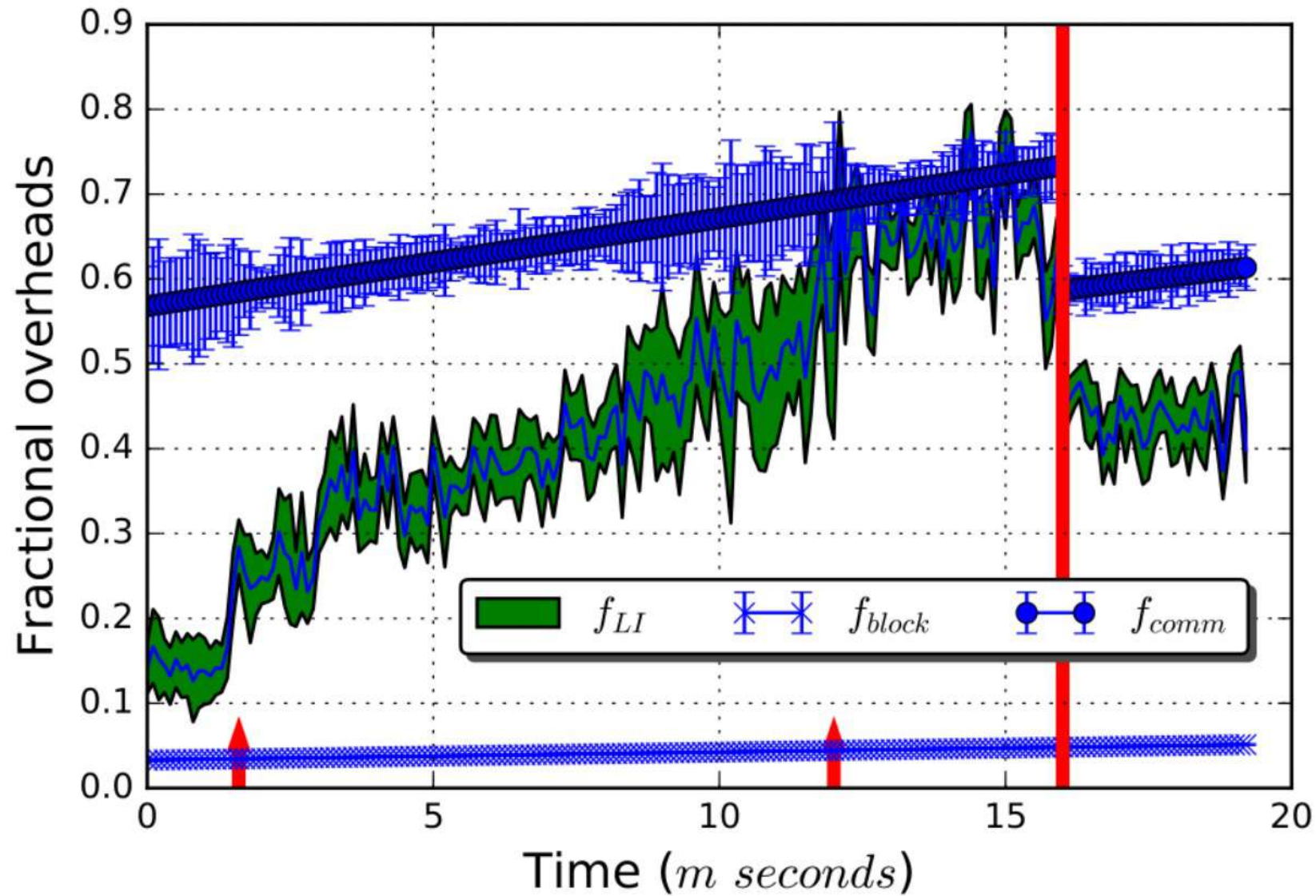


How does it look in the cell simulation?

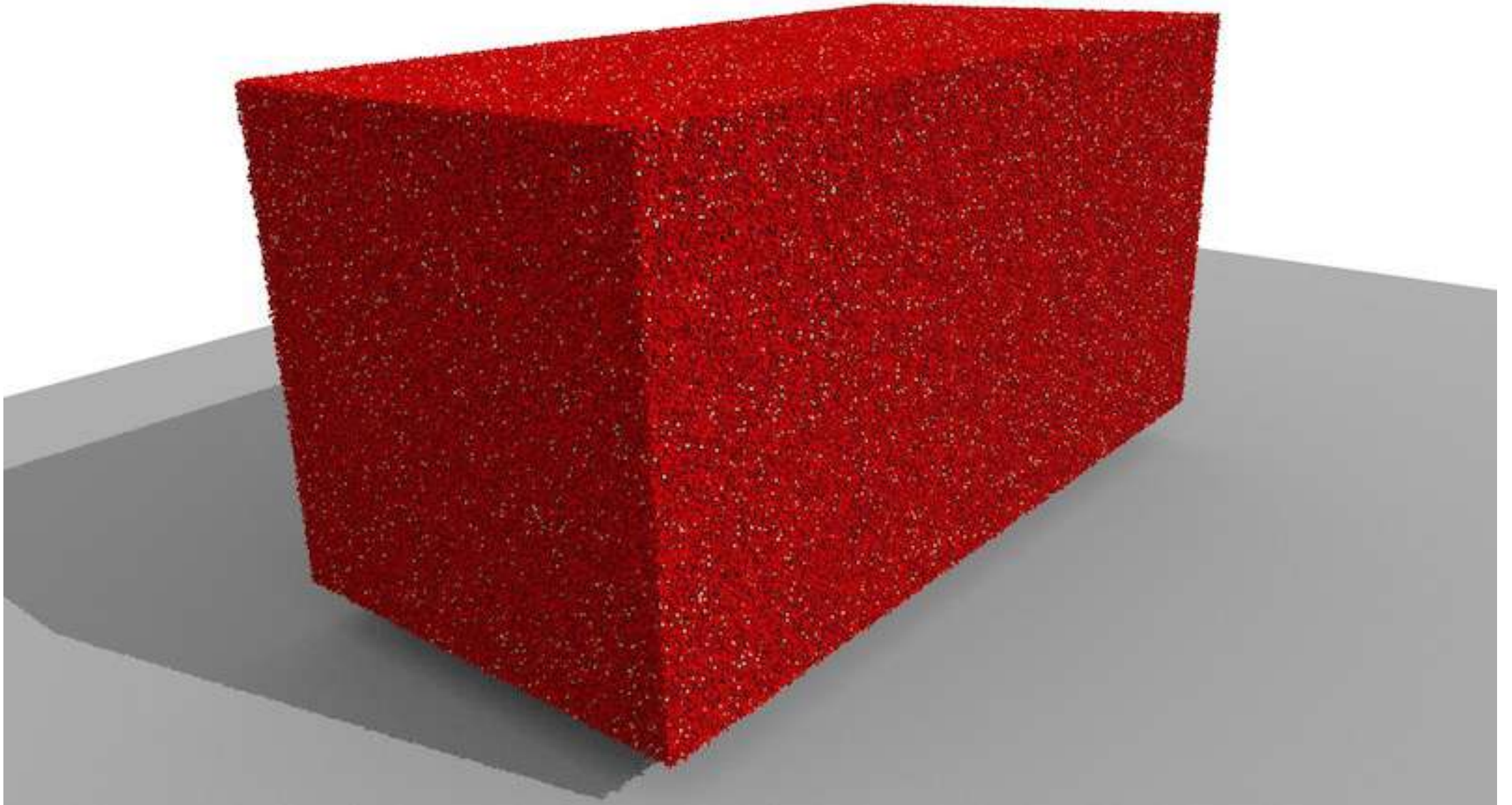
- Initialise with N atomic blocks
 - Here $N = 80$
 - Minimal size of atomic block determined by max RBC size
 - We use Palabos functionality to do this
- Load balance over available processors, we give atomic blocks to processors
 - Metis, using weights on the nodes, applying the multilevel KL.
 - Here $p = 12$
- Merge atomic blocks when possible (when in the same processor) to save on “communication”
- Then, at run time the load balance changes, so we rebalance every 2000 timesteps



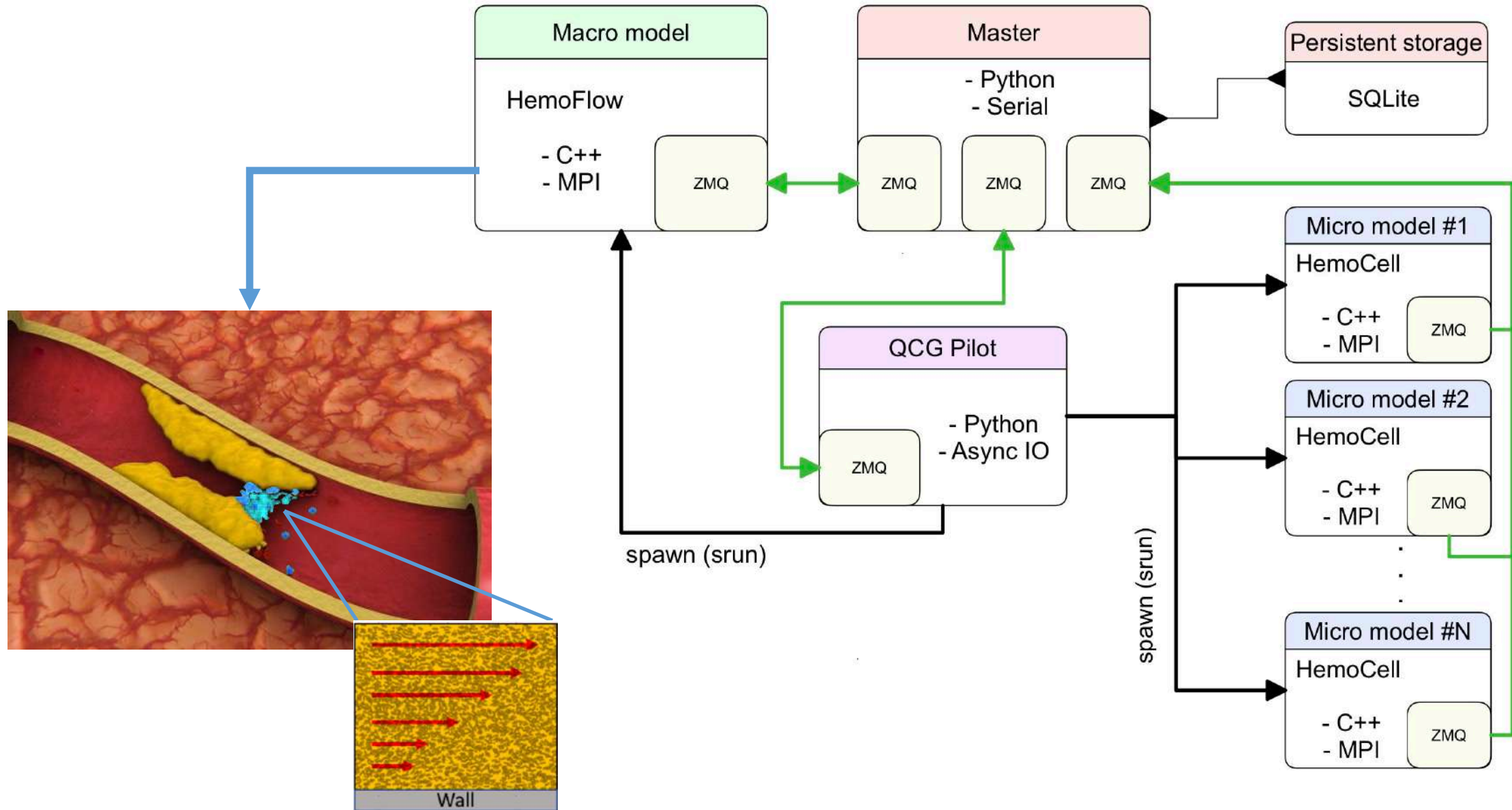
Effects of load-balancing in a large-scale simulation



These techniques make larger domains and more details possible



Multi-scale coupled models (in development)



Team, friends & partners in development

