Efficient simulations of patient-specific electrical heart activity on the DGX-2

simula

eX3 Experimental Infrastructure for Exploration of Exascale Computing

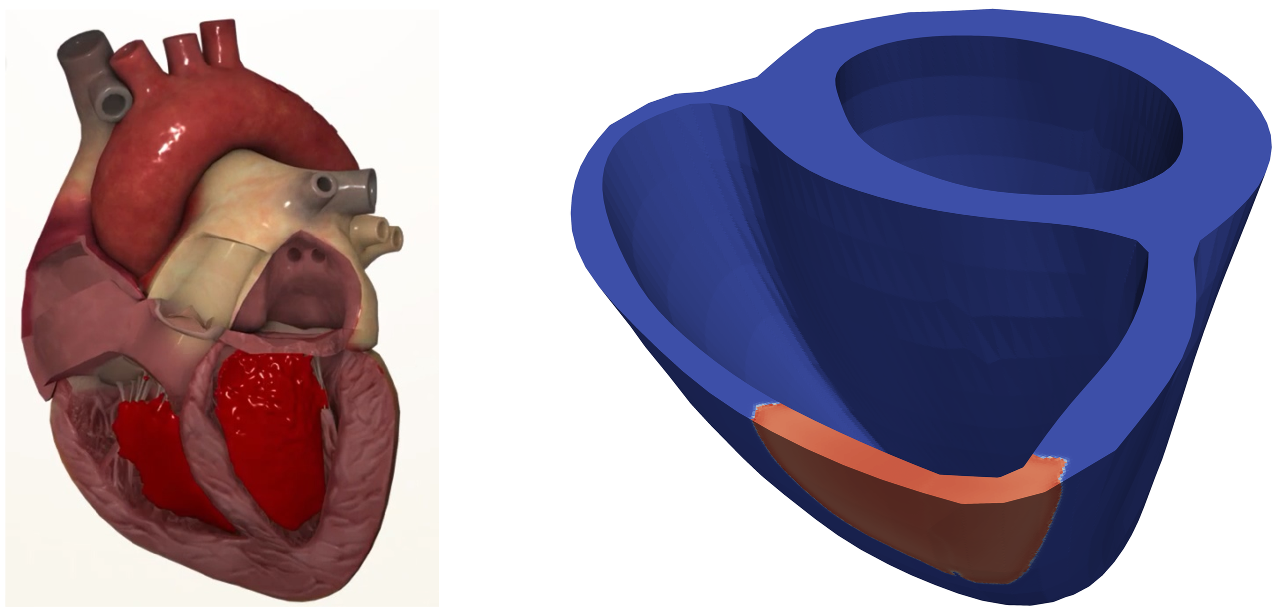
Kristian Gregorius Hustad [1,2] Xing Cai [1,2] Johannes Langguth [1] Hermenegild Arevalo [1,3]

[1]Simula Research Laboratory, Norway [2]Department of Informatics, University of Oslo [3]Center for Cardiological Innovation, Oslo

CENTER FOR CARDIOLOGICAL INNOVATION

Introduction

Cardiovascular disease is the **leading cause of death** in the industrialised world. Patients who have suffered a heart attack have an **elevated risk of developing arrhythmia**. Computer simulations of the electrical activity in the heart can be used **predict the risk of arrhythmia** in these patients.



(a) Illustration of the heart (b) Tetrahedral mesh for the heart ventricles

Computational problem

Reaction-diffusion problem with two kernels that run at every time step.

Reaction kernel

- Solves system of ordinary differential equations (ODEs) $\frac{\partial v}{\partial t} = -I_{\text{ion}}(v, s)$ describing what happens **within each cell**
- Lots of expensive exponential function evaluations
- Large memory traffic from updating 19 FP64 values describing the state of each cell ($2 \cdot 19 \cdot 8\,\text{B} = 304\,\text{B}$)
- High register usage (128) $\implies$ occupancy ≤ 0.25

Diffusion kernel

- Solves the diffusion equation $\frac{\partial v}{\partial t} = \frac{\lambda}{1+\lambda}\nabla \cdot (\hat{M}_i \nabla v)$ describing how the signal spreads **between cells**.
- Discretised with explicit finite volume method for **unstructured tetrahedral meshes**
  - Uses the 4 first-order and the 12 second-order neighbours.
  - $\implies$ performs a single **sparse matrix-vector multiplication (SpMV)**
- Heavily memory bound. Need communication between (GPUs owning) neighbouring partitions.

Total cost

- $\Delta t = 20\,\mu\text{s} \implies 50\,000$ time steps per heartbeat (at 60 bpm).
- We use patient-specific meshes with **6–15 million cells**.
- Minimum memory traffic per cell step $\tau^{\text{min}}_{\text{cell step}} = 520$ bytes
- Minimum memory traffic per heartbeat $\tau^{\text{min}} \geq$ time steps $\cdot N \cdot \tau$ $\implies$ **156–390 TB of memory traffic** per heart second.
  - Aggregate memory bandwidth of DGX-2: $16 \cdot 887\,\text{GB/s} \approx 14.2\,\text{TB/s}$
  - $\implies$ Lower bound on time (using 16 GPUs): **11.0–27.5 seconds**

Impact

Simulations on patient-specific heart models could provide doctors with not only **safer** but also **more accurate** results than current invasive procedures permit.

Parallelisation

Find the subset of the partition $\mathcal{P}_i$ that is needed by other partitions, and label this the **partition separator**. Compute the separator first, then start sending it while computing the remaining values (the **partition interior**).
Create one partition per GPU. Use the CPUs purely for orchestrating the computation (this way we avoid CPU-GPU data transfers during the computation).

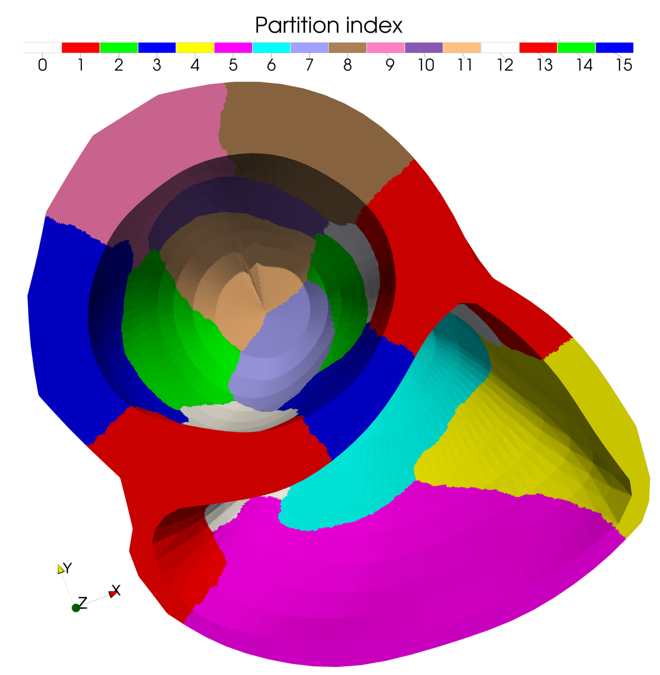


Figure 2. Mesh with 16 partitions

The NVSwitches in the DGX-2 enable **low-latency, high-bandwidth communication** between any pair of GPUs, ensuring that we don't hit any communication bottlenecks even when using all 16 GPUs.

Optimisation

Reaction kernel

Reduce number of floating point operations by

- exploiting mathematical identities to reduce the number of exponential function evaluations
- hand-optimising the kernel for improved re-use of expressions that have already been computed

Diffusion kernel

Kernel consists of a single SpMV, $\mathbf{v}^{n+1} = \mathbf{Z}\mathbf{v}^n$. $\mathbf{Z}$ has **at most 16** off-diagonal non-zero elements per row. Diagonal is stored in dense array. Off-diagonal elements are stored in ELLPACK format. Transpose 32×16 blocks for coalesced memory accesses within each warp.

Reorder matrix for better caching. Use the METIS graph partitioner on the connectivity graph for $\mathbf{Z}$ to create many small clusters, each small enough to fit in L1/L2 cache. Measured memory traffic is $\sim 1\%$ greater than theoretical minimum.

Challenges

- Cell model kernel is more expensive for non-excited cells $\implies$ dynamic load imbalance
- The problem size per GPU becomes small with 16 GPUs
  - One thread per cell
  - Full occupancy on a V100 is $80 \cdot 2048 = 163\,840$ threads.

Multi-GPU scaling on the DGX-2

GPUs	Time (s)	Scaling efficiency	Ratio of theoretical max performance
1	400.10	1.000	0.874
2	208.50	0.959	0.838
4	105.57	0.947	0.828
8	53.80	0.930	0.812
16	28.16	0.888	0.776

Table 1. Time to simulate a heartbeat (1 s of heart activity). $N = 11\,688\,851$, and the number of time steps is 50 000.

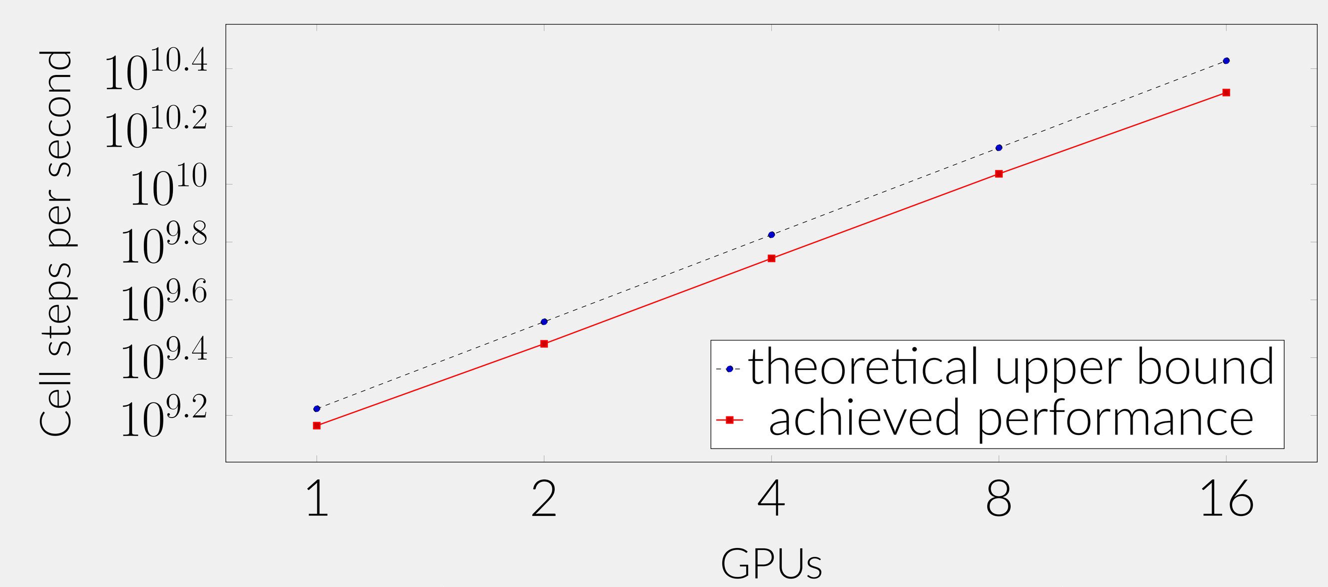
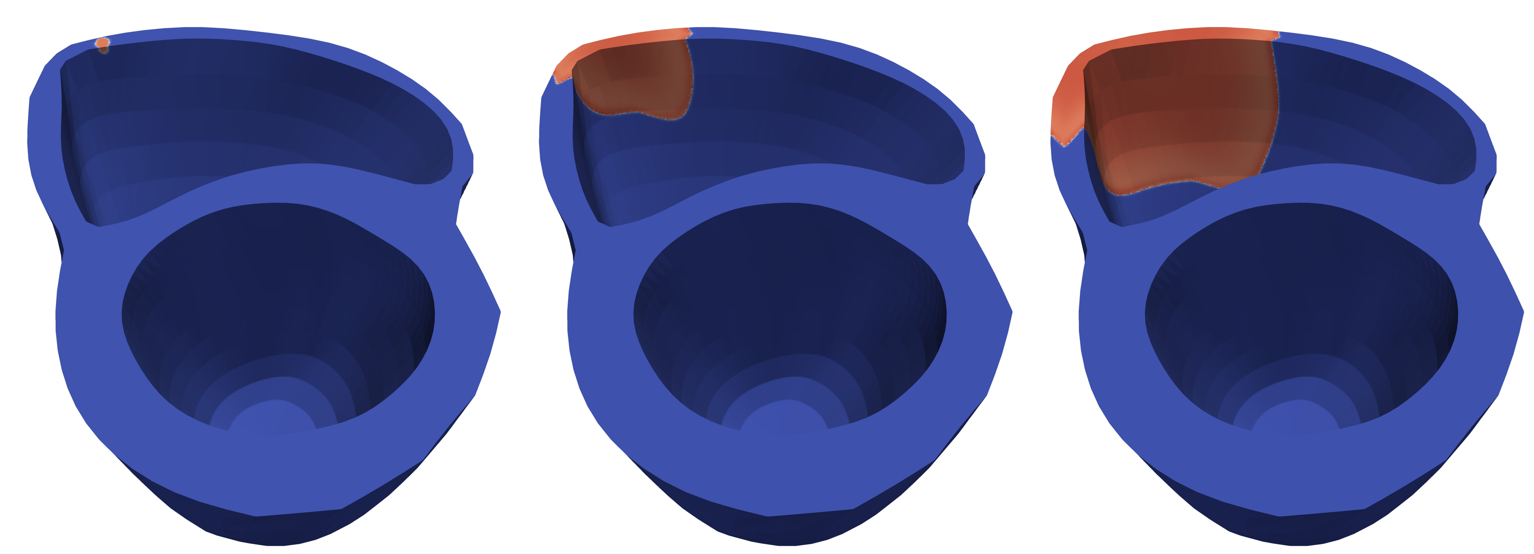


Figure 3. Throughput (measured in cell steps per second) vs # of GPUs.

Using all 16 GPUs in the DGX-2, we are able to run the simulation at $\geq \frac{1}{30}$ of real-time. Assuming a heart rate of 60 bpm, we achieve a performance of **2 heartbeats per wall clock minute**.

Simulation results



(a) $t = 2$ ms (b) $t = 50$ ms (c) $t = 100$ ms

Acknowledgements

The computations in this poster were performed on equipment provided by the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under contract 270053.

References

[1] Hermenegild J. Arevalo, Fijoy Vadakkumpadan, Eliseo Guallar, Alexander Jebb, Peter Malamas, Katherine C. Wu, and Natalia A. Trayanova.
Arrhythmia risk stratification of patients after myocardial infarction using personalized heart models.
Nature Communications, 7:11437, 05 2016.

[2] Kristian Gregorius Hustad.
Solving the monodomain model efficiently on GPUs.
MSc thesis, Department of Informatics, University of Oslo, 2019.

[3] Johannes Langguth, Mohammed Sourouri, Glenn Terje Lines, Scott Baden, and Xing Cai.
Scalable heterogeneous CPU-GPU computations for unstructured tetrahedral meshes.
IEEE Micro, 35:6–15, 07 2015.