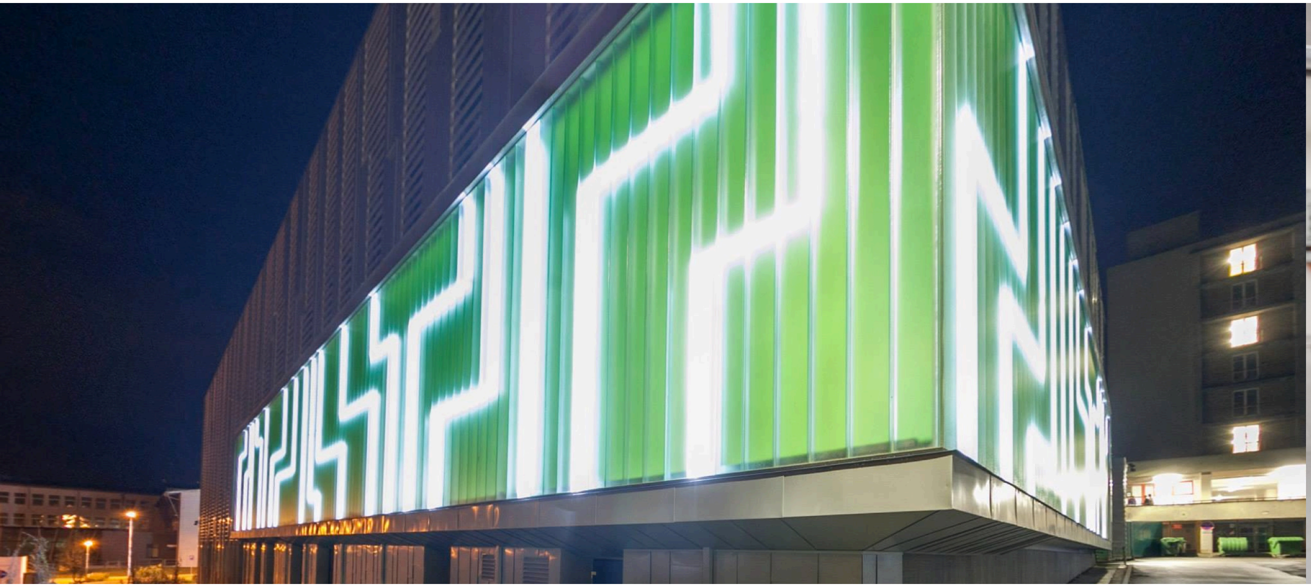
IT4Innovations national supercomputing center



Selected Intel Xeon Phi Accelerated Libraries developed at IT4Innovations

Lubomír Říha ESPRESO Library
Michal Merta BEM4I Library
Milan Jaroš CyclesPhi Render

Approaches to Accelerate Application using Xeon Phi (KNC)

IT4Innovations national01$#&0 supercomputing center@#01%101

ESPRESO

- Acceleration of the FETI solver
- FEM produces sparse matrices – new method converts sparse matrices to dense
- Memory bound code – utilizes high bandwidth of the KNC memory

BEM4I

- Acceleration of the BEM matrix assembler
- Compute bound code – relies on high performance (wide) vector units of the KNC core

CyclesPhi

- Parallelization of the complex third party code using OpenMP and Offload pragmas to utilize KNC
- Developed MPI parallelization to support HPC environment
  - Enables real-time raytracing in interactive mode

ESPRESO Library

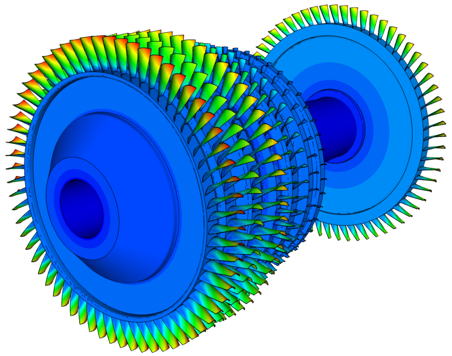
Library based on FEM/BEM with Massively parallel sparse linear solver designed to take full advantage of todays most powerful peta-scale supercomputers

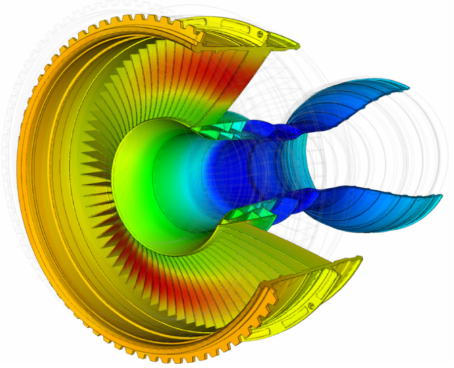
Preprocessing
- ANSYS
- ESPRESO Generator
- **ESPRESO API**
- **OpenFOAM**

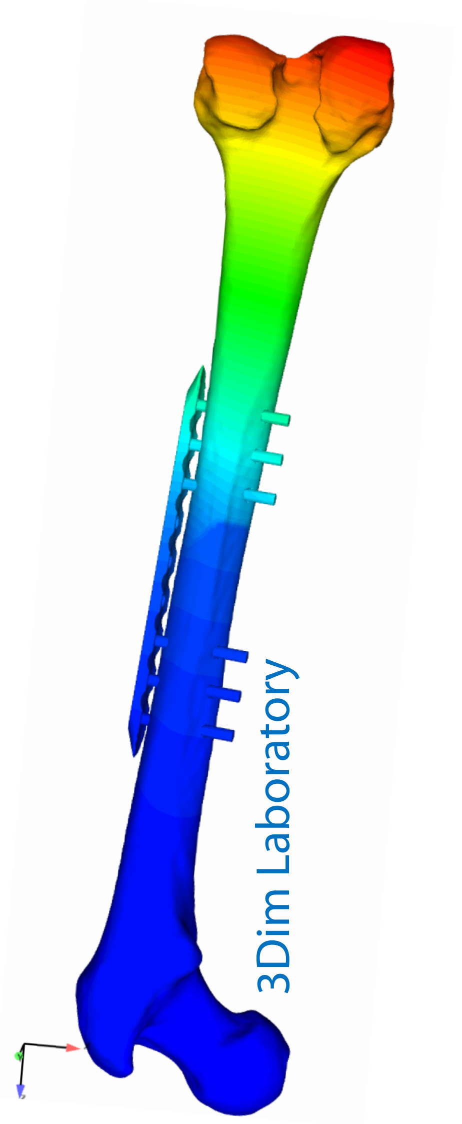
ESPRESO C++ library
- Mesh Processing
- Matrix Assembler FEM/BEM (BEM4I)
- TFETI/Hybrid TFETI Solvers
- CPU | KNC, KNL | GPU

Postprocessing
- Visualization
- EnSight / Visit
- Paraview Catalyst

intel Parallel Computing Center







ESPRESO and Large Problems

4th in TOP500 LIST

18,688 AMD Opteron 6274 16-core CPUs
18,688 Nvidia Tesla K20X GPUs

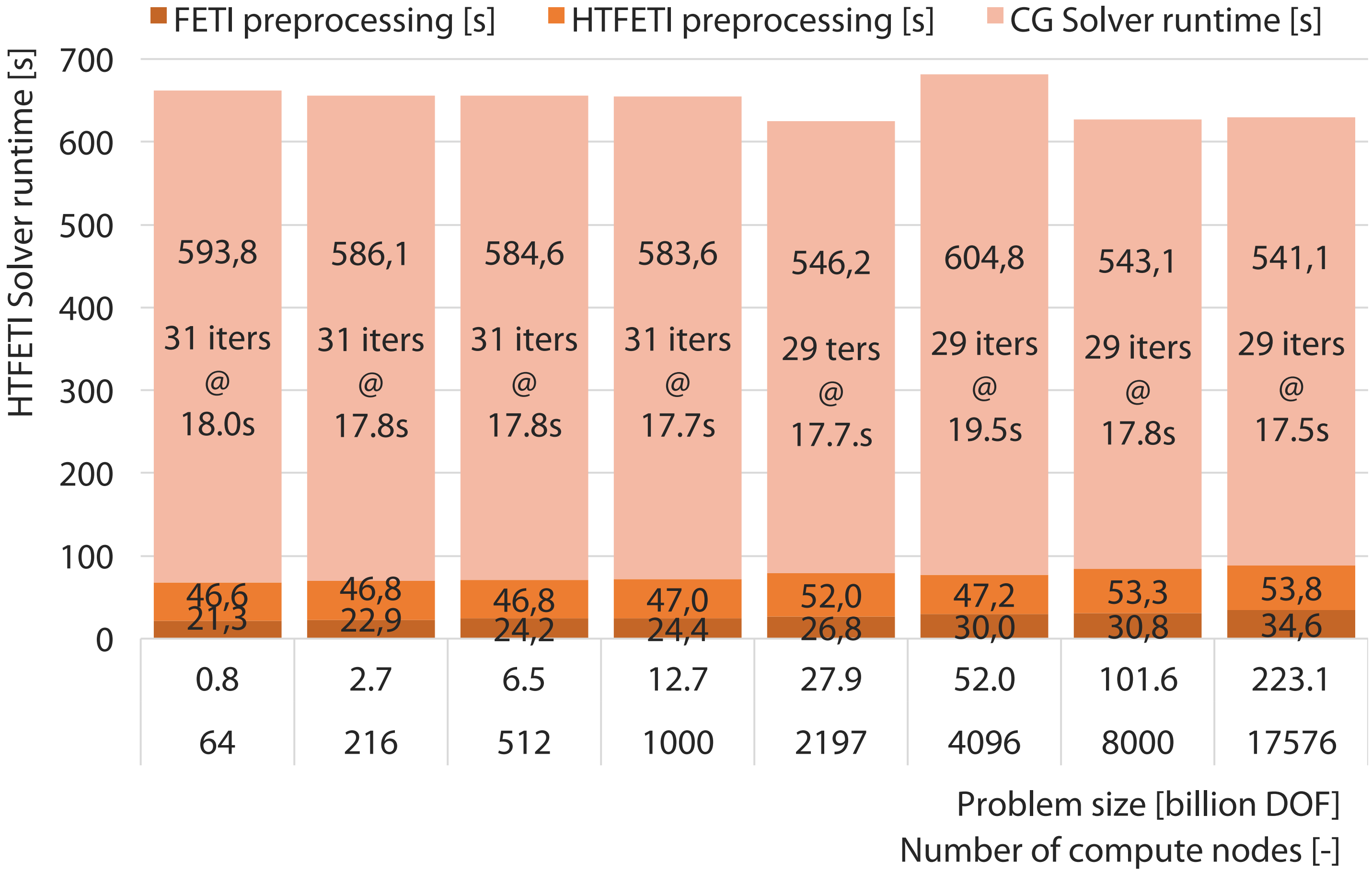
2.7 million core hours dedicated to:

- scalability optimization of ESPRESO
- optimization of GPU accelerated version for large scale problems

Weak Scalability Test

Up to **223 billion** DOF on 17576 Compute Nodes (281 216 cores)

Heat transfer

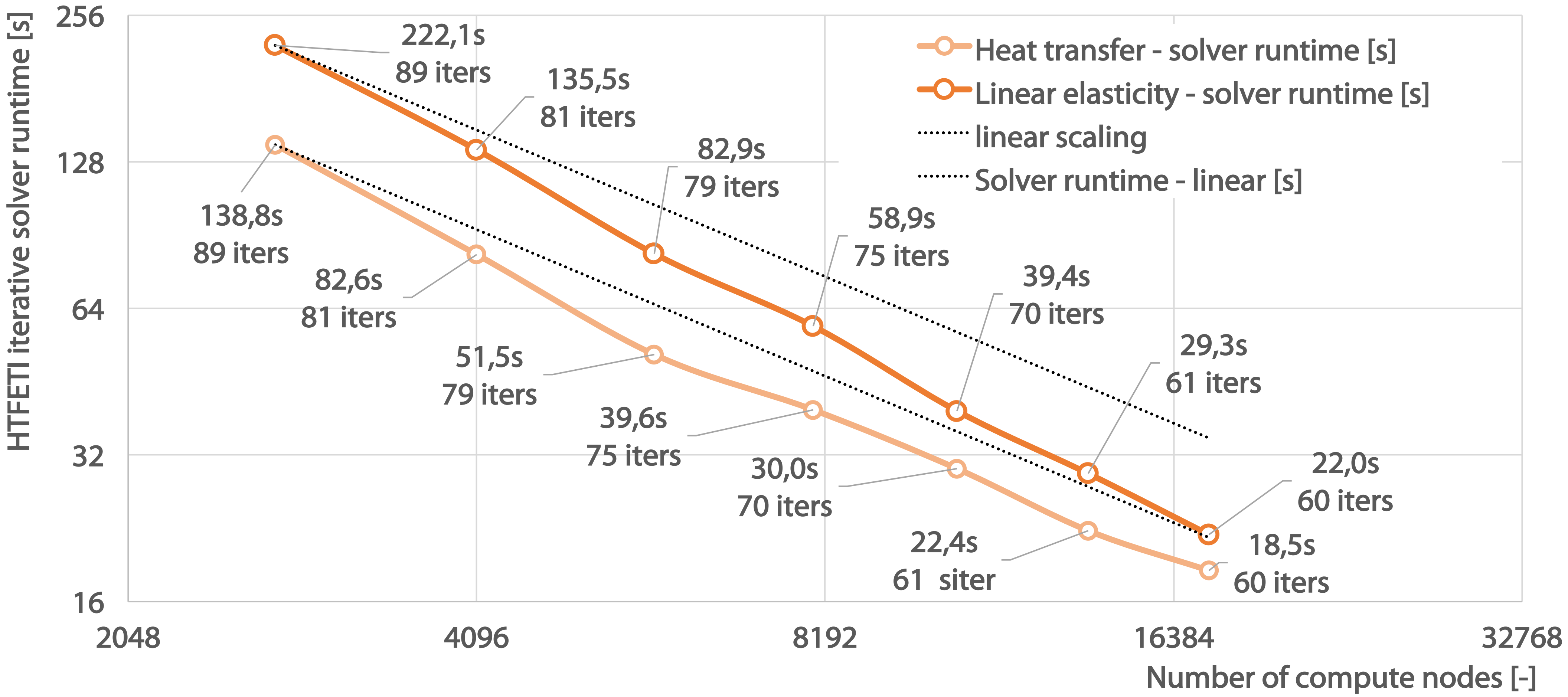


Scalability of ESPRESO

IT4Innovations national supercomputing center

Heat transfer 20 billion DOF on up to 17 576 Compute Nodes (281 216 cores)
Linear elasticity 11 billion DOF

ORNL Titan 4th in TOP500 LIST



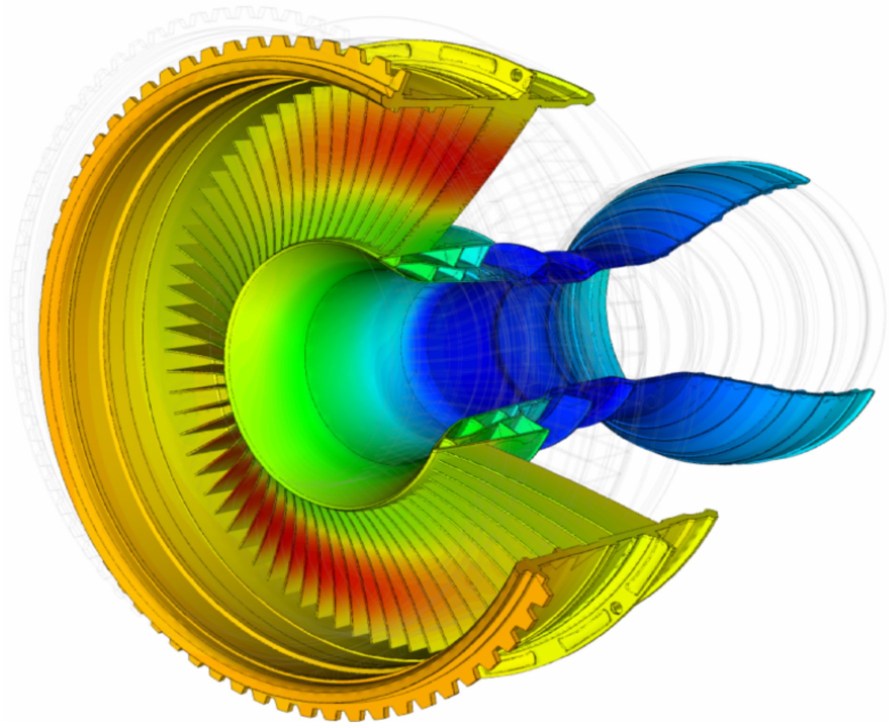
IT4Innovations national supercomputing center

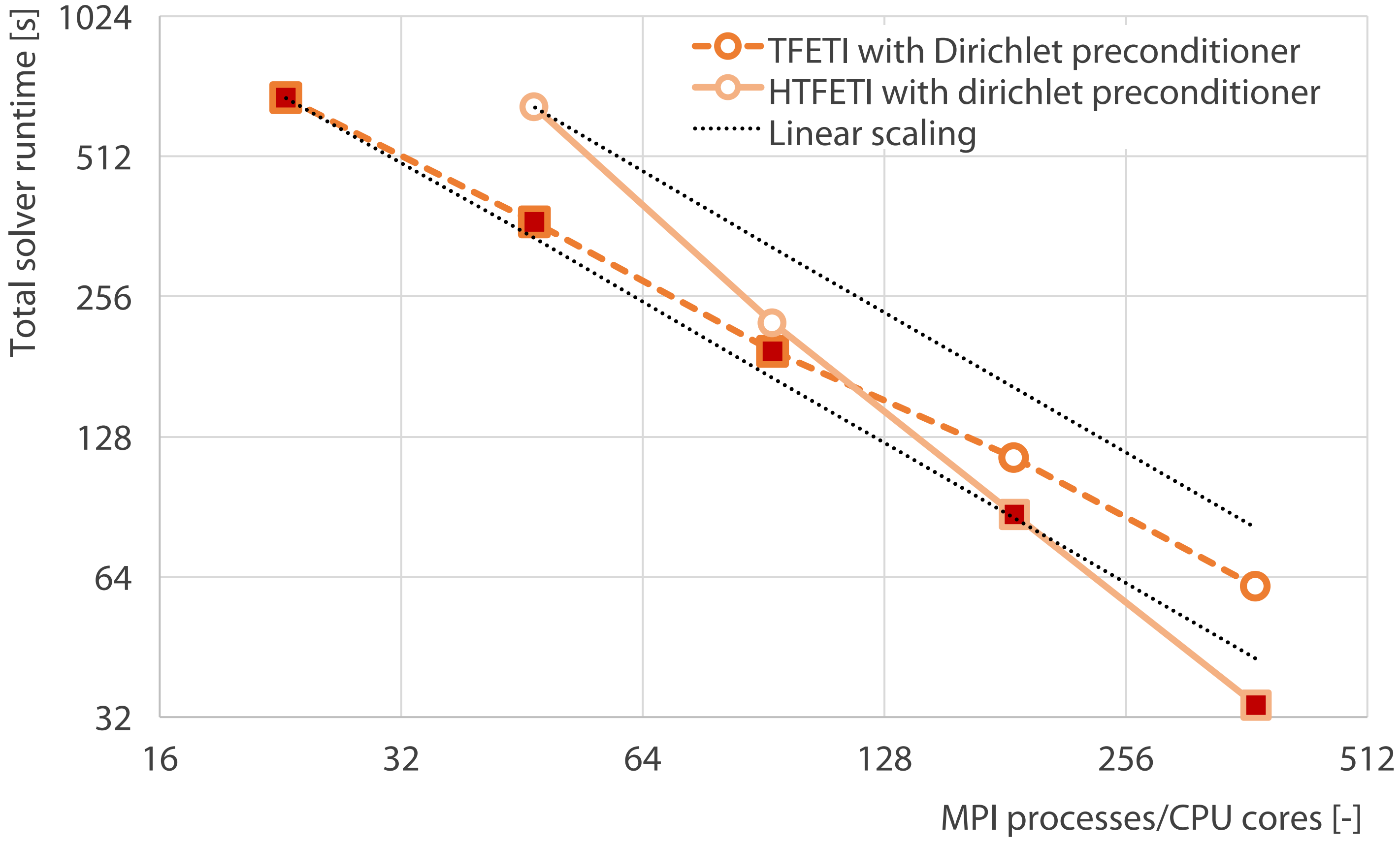
Scalability for Real World Problems

300 million unknown - ANSYS Workbench real world problem

Linear elasticity – Hybrid FETI with Dirichlet preconditioner

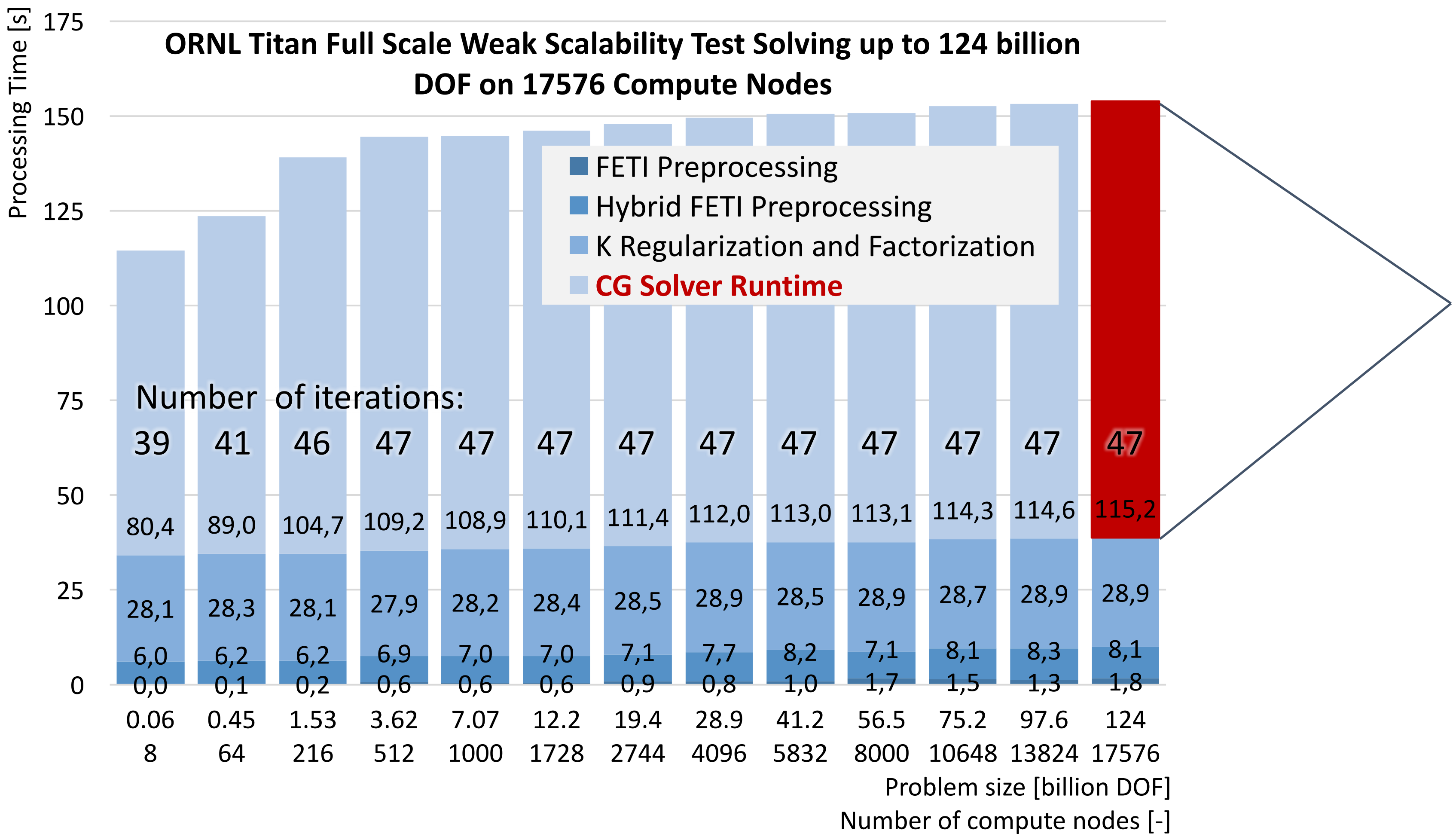
IT4Innovations – SALOMON Supercomputer – only Haswell CPUs are used here





Experiment setup:
Dirichlet preconditioner
regular CG solver
HTFETI method

Motivation for Xeon Phi Acceleration



This part of the solver is accelerated by the Intel Xeon Phi co-processor

IT4Innovations national supercomputing center

Hardware Acceleration of FETI solvers

Local Schur Complement (LSC) Method

$r_0 := b - Ax_0;\ u_0 := M^{-1}r_0;\ p_0 := u_0$
for $i = 0, \ldots, m-1$ **do**
$\quad s := Ap_i$
$\quad \alpha := \langle r_i, u_i\rangle \,/\, \langle s, p_i\rangle$
$\quad x_{i+1} := x_i + \alpha p_i$
$\quad r_{i+1} := r_i - \alpha s$
$\quad u_{i+1} := M^{-1} r_{i+1}$
$\quad \beta := \langle r_{i+1}, u_{i+1}\rangle \,/\, \langle r_i, u_i\rangle$
$\quad p_{i+1} := u_{i+1} + \beta p_i$
end for

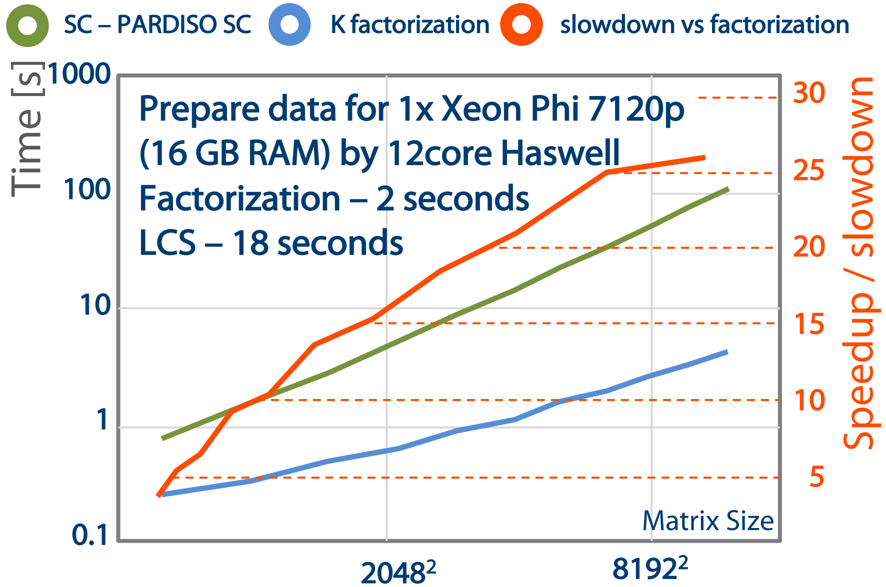
Pre-processing – K factorization

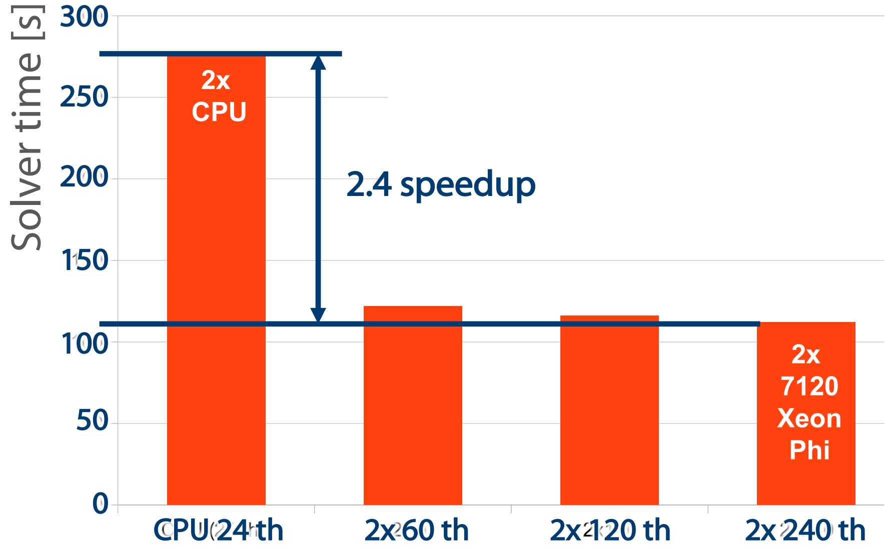
1.) $x = B_1^T \cdot \lambda$ - **SpMV**
2.) $y = K^{-1} \cdot x$ - **solve**
3.) $\lambda = B_1 \cdot y$ - **SpMV**
4.) stencil data exchange in λ
- MPI – Send and Recv
- OpenMP – shared mem. vec

Pre-processing - $S_c = B_1 K^{-1} B_1^T$ → MIC

1.) λ → MIC - **PCIe transfer from CPU**
2.) $\lambda = S_c \cdot \lambda$ - **DGEMV, DSYMV on MIC**
3.) λ ← MIC - **PCIe transfer to CPU**
4.) stencil data exchange in λ
- MPI – Send and Recv
- OpenMP – shared mem. vec

Using Schur complement we are still memory bounded, but fast Intel Xeon Phi memory can be now fully utilized due usage of dense matrices. This is the main factor that brings the speedup when compared to CPU which is two generations ahead.





Architecture Comparison for LSC Method

	Number of cores	Peak floating point performance SP/DP [GFLOPS]	Theoretical Memory Bandwidth [GB/s]	Memory type
Intel Xeon E5-2680v3	12	960 / 480	68	DDR4
Intel Xeon Phi 7120p	61	2420 / 1210	~~352~~ **180**	GDDR5
Intel Xeon Phi 7210	64	5325 / 2662	102 / 400	DDR4/MCDRAM
NVIDIA Tesla K80 (1 chip)	1597 SP / 533 DP	4365 / 1455	240	GDDR5
NVIDIA Tesla K20X (Titan)	2688 SP / 896 DP	3950 / 1310	250	GDDR5
NVIDIA Tesla P100 (PCIe)	3584 SP / 1792 DP	8345 / 4217	720	HBM2

Method	Problem size and decomposition		Solve time[s]	Speedup by LSC method for solve				
	number of subdomains [-]	Subdomain size [DOF]	PARDISO CPU only	**CPU only**	**KNC & CPU***	**KNL only**	**K80 & CPU***	**P100 & CPU***
	Heat Transfer problem							
TFETI	512	6859	21.4	1.3	**2.1**	2.2	1.5	4.0
HTFETI	512	6859	26.0	1.2	**2.4**	2.5	1.6	3.5
	Linear elasticity problem							
TFETI	512	3993	20.4	1.1	**1.8**	2.5	1.3	3.5
HTFETI	512	3993	21.2	0.9	**1.8**	2.4	1.2	2.7

*** Note:** CPU does not process LSCs

Large Scale Tests on Salomon

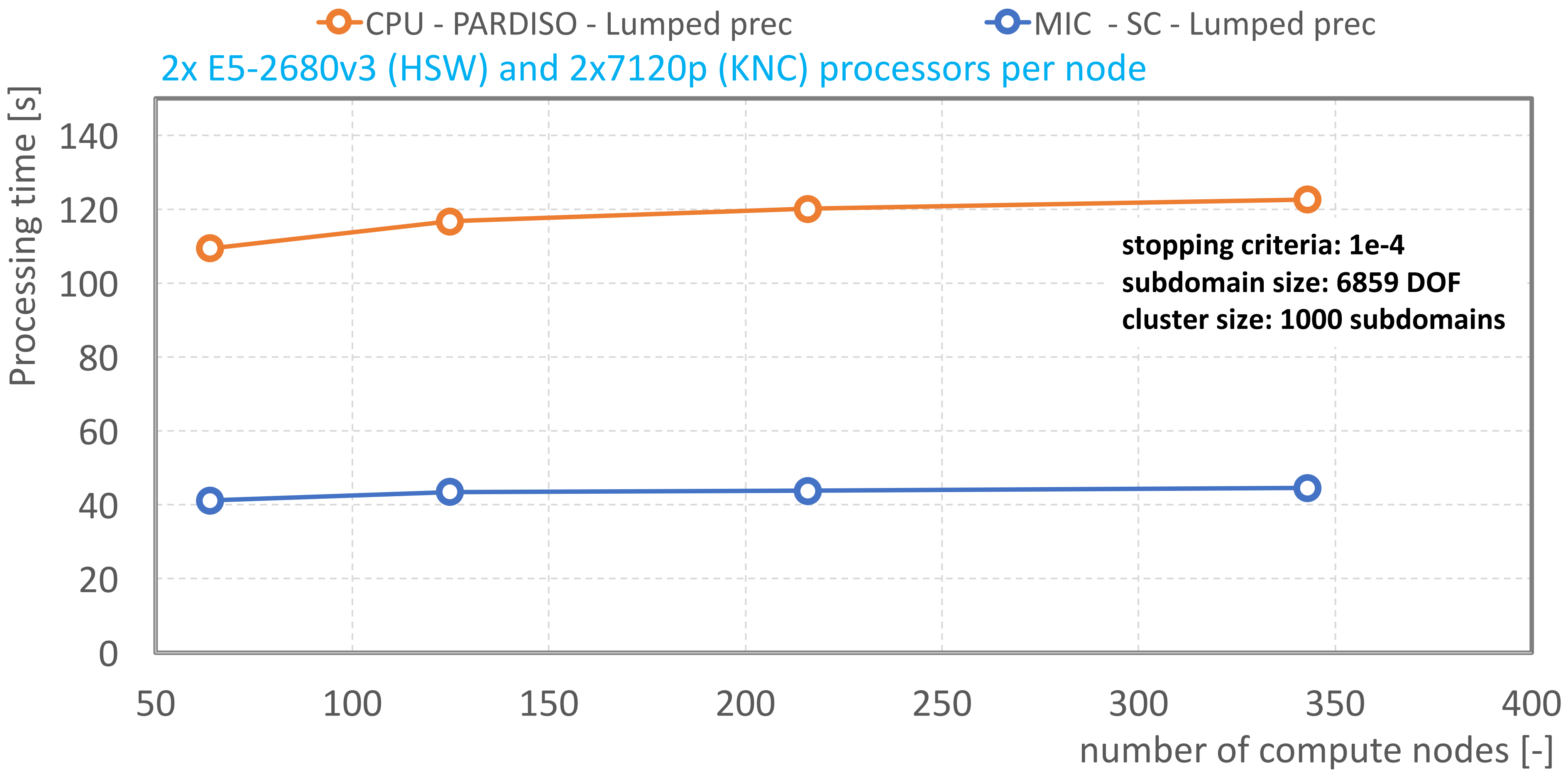
330 – 1776 million DOF Hybrid FETI CG Solver Runtime

Heat Transfer – CG Solver Runtime w. Lumped Prec.

IT4Innovations Salomon Supercomputer

Speedup

2.4



Solving Real World Problems – Challenges

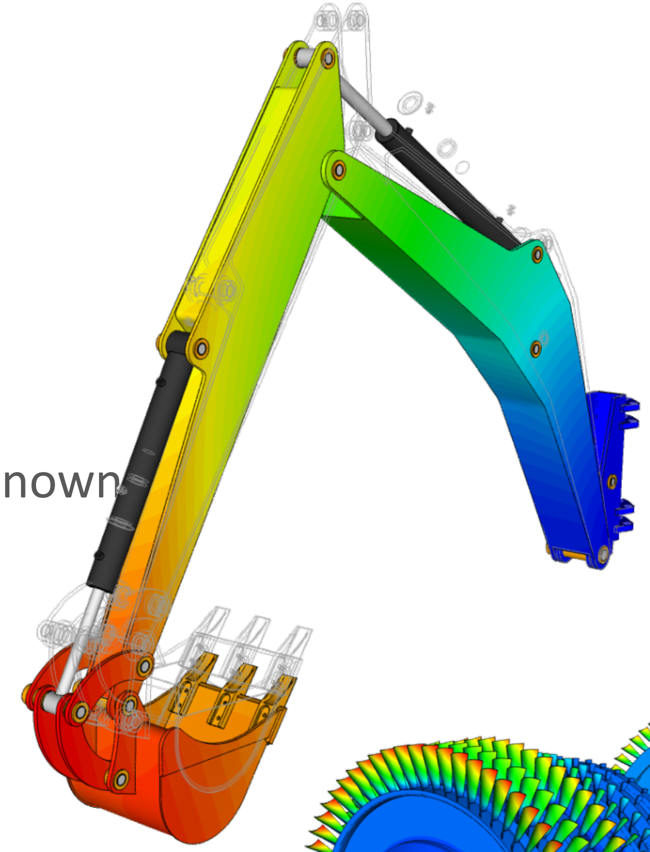
Input data in from **ELMER**, **OpenFOAM** or **Ansys Workbench**

Worse convergence – requires high number of iterations

- **requires efficient Dirichlet preconditioner**
- Full orthogonal CG solver with restarts for large coefficient jumps
- HTFETI method with clusterization by local kernels

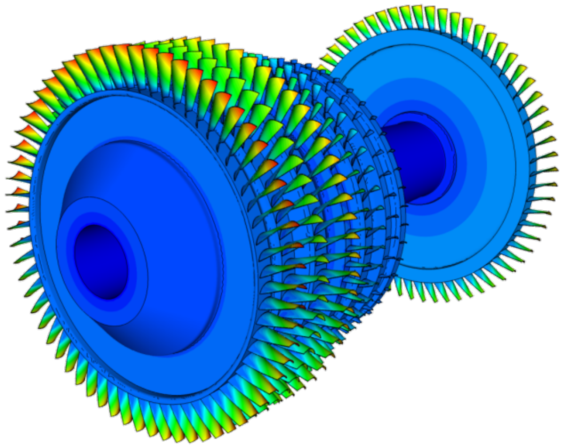
Specific requirements for "smaller problems" up to 500 million unknowns

- **multiple MPI processes share single node**
  - **resource sharing control for accelerators**
- **efficient use of all resources**



Required enhancements in Xeon Phi Acceleration

1. **Acceleration of Dirichlet preconditioner**
2. **Load balancing between CPU and Xeon Phi**
3. **Efficient sharing of one Xeon Phi by multiple MPI processes**



IT4Innovations
national
supercomputing
center

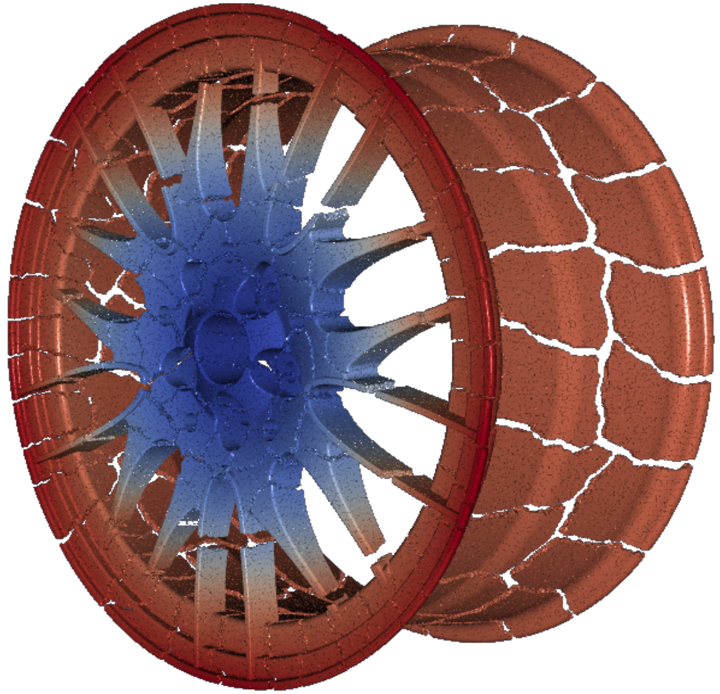
Solving Real World Problems on Xeon Phi

1.) Acceleration of the Dirichlet Preconditioner

	2 MPI per MIC		
	CPU	MIC	Speedup
MKL PARDISO / LSC action time [s]	0.179	0.087	2.1
Dirichlet preconditioner action time [s]	0.105	0.060	1.8

Tire Rim benchmark with Dirichlet Preconditioner

Linear elasticity – 120 million unknown



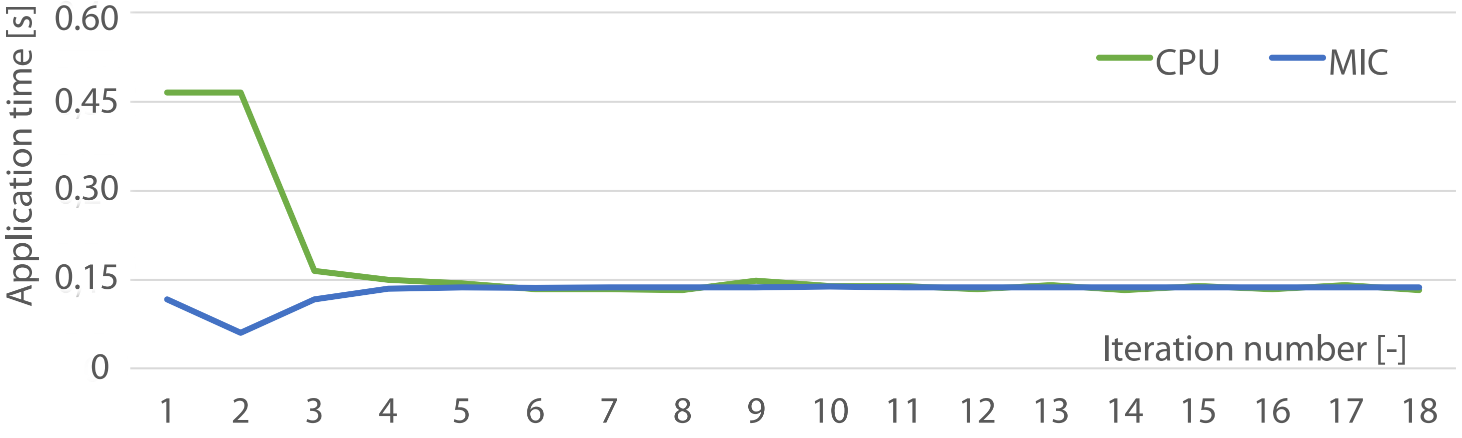
64 Salomon's compute nodes

128 7120p accelerators

~2 million unknowns per node

2.) Dynamic Load balancing between CPU and accelerator

Minimizes runtime by automatically splitting the workload



3.) Efficient sharing of one Xeon Phi by multiple MPI processes

2 MPI ranks per Xeon Phi

	CPU with PARDISO solver			MIC with LSC method			
# of subdomains per MPI rank [-]	Iteration time [s]	CG solver runtime [s]	number of iterations [-]	Iteration time [s]	CG solver runtime [s]	number of iterations [-]	speedup [-]
128	0.298	**88.0**	**295**	0.146	**42.9**	**294**	**2.0**

Transient Problem on Xeon Phi

Cube benchmark

Heat Transfer – 5 million unknowns per node

A good use case for LSC and Xeon Phi acceleration

1x Salomon's compute node
2x 7120p accelerators
~2.0 speedup which includes LSC and no. prec.

Benchmark execution:

```
mpirun -n 2 ./espreso -c espreso-phi-bench.ecf 0  1 1 2  6 6 10  19 19 19
```

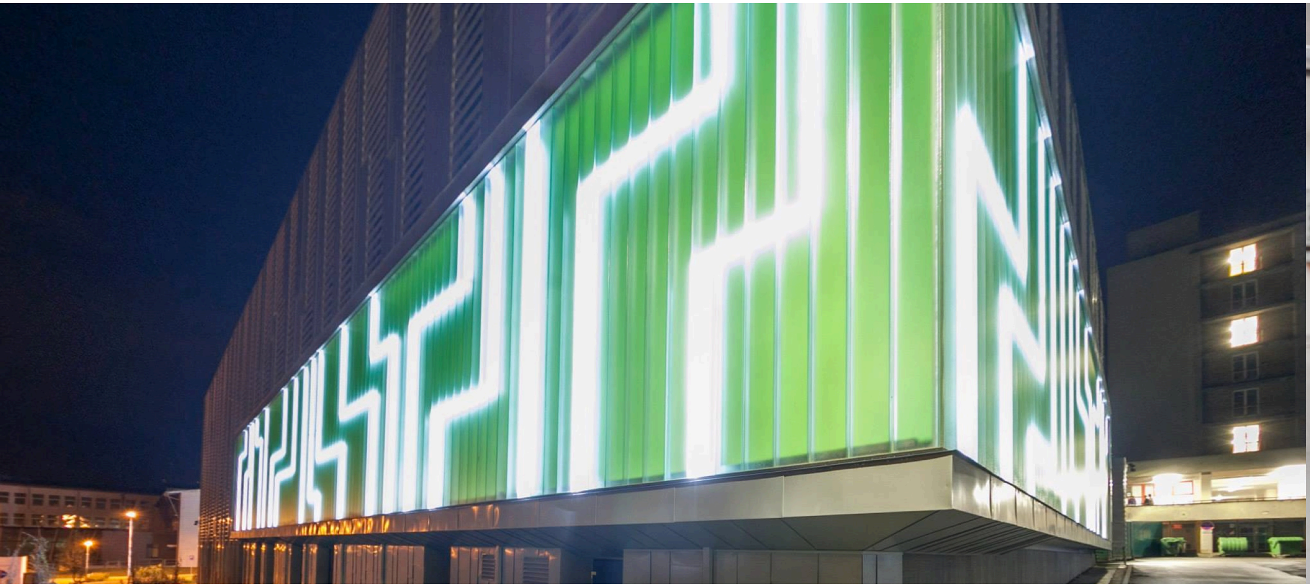
Problem statistics:

```
Generate grid of hexahedrons
Coordinates loaded - total number of nodes: 5 051 950
Elements loaded - total number of elements: 4 938 480
```

Time step	Transient solver stage	MIC LSC time [s]	CPU PARDISO time [s]	speedup [-]	savings [s]
preprocessing	FETI Preprocessing	0.01	0.01	1.00	0.0
preprocessing	Schur Complement asm.	54.8	0.00	---	-54.8
preprocessing	K factorization	0.00	2.9	---	2.9
0	CG Solver runtime	17.8	35.6	2.00	17.8
1	RHS update	0.01	0.01	1.0	0.0
1	CG Solver runtime	17.2	34.7	2.02	17.5
2	RHS update	0.01	0.01	1.0	0.0
2	CG Solver runtime	16.4	34.0	2.07	17.6
3	RHS update	0.01	0.01	1.0	0.0
3	CG Solver runtime	16.1	33.5	2.08	17.4
4	RHS update	0.01	0.01	1.0	0.0
4	CG Solver runtime	16.2	33.8	2.08	17.6
9	RHS update	0.01	0.01	1.0	0.0
9	CG Solver runtime	16.2	32.8	2.03	16.7
...					

10 time steps	Total time	219s	339s	1.5	120s
100 time steps	Total time	1694s	3402s	2.0	1708s

IT4Innovations
national
supercomputing
center



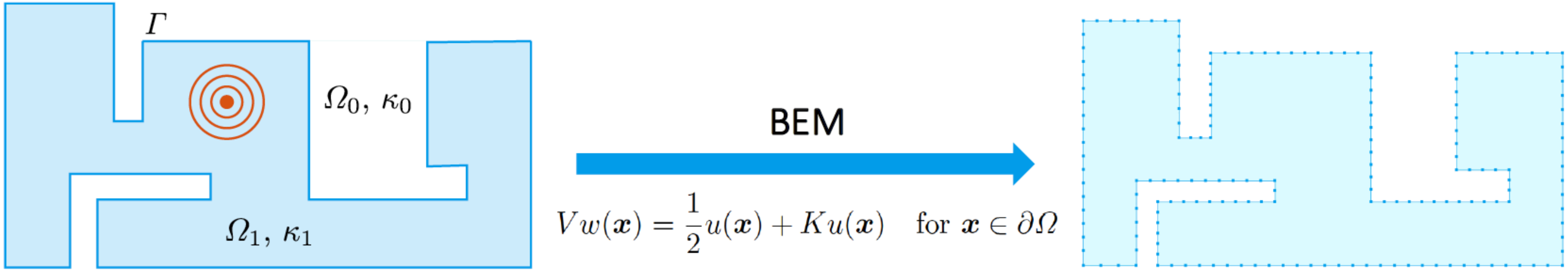
BEM4I Library

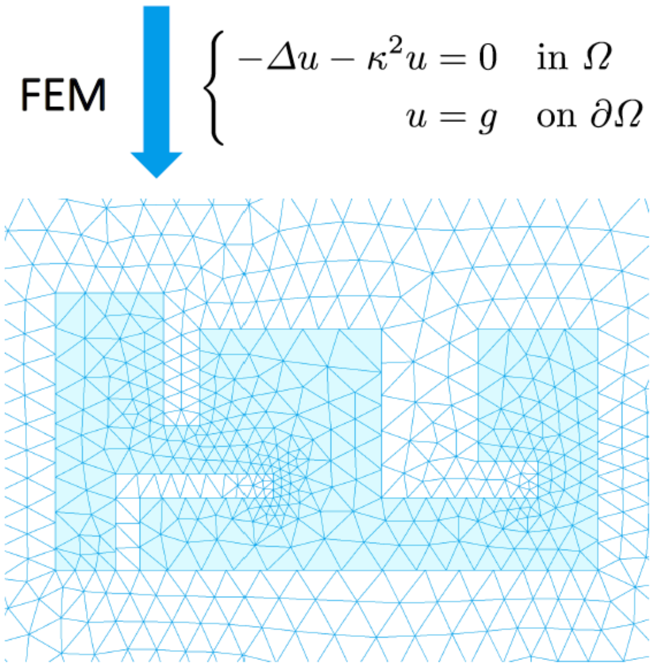
Michal Merta

BEM4I library

IT4Innovations national supercomputing center

Parallel boundary element library



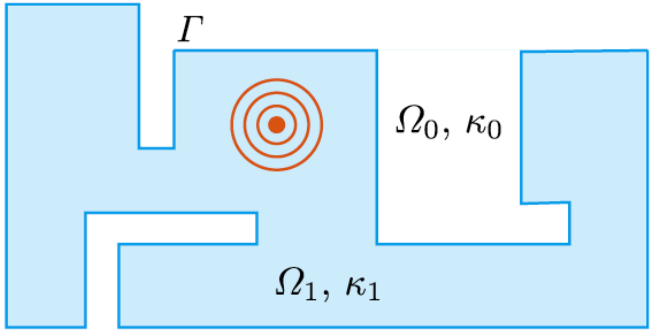


- 3D boundary element solver in C++
- Laplace, Helmholtz, Lamé, wave equations
- SIMD, OpenMP, MPI
- Offload to Intel Xeon Phi
- BEM4+ESPRESO = {BETI, FETI/BETI coupling}

BEM4I library

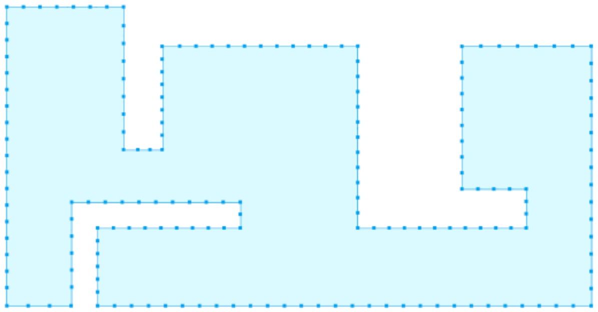
IT4Innovations
national
supercomputing
center

Parallel boundary element library



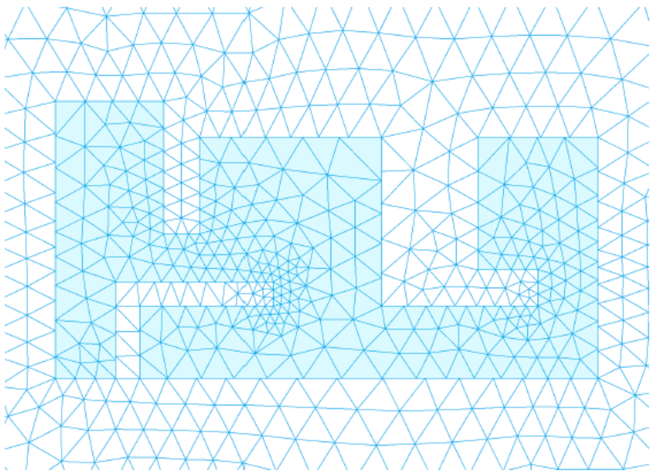
BEM

$$Vw(\boldsymbol{x}) = \frac{1}{2}u(\boldsymbol{x}) + Ku(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \partial\Omega$$



FEM

$$\begin{cases} -\Delta u - \kappa^2 u = 0 & \text{in } \Omega \\ u = g & \text{on } \partial\Omega \end{cases}$$



Laplace	Lamé	Helmholtz	Wave

Integrators and matrix assemblers

Common core: matrices, vectors, solvers

Wrappers to BLAS, LAPACK, Eigen, Metis

BEM4I library

Porting BEM on modern Intel architectures

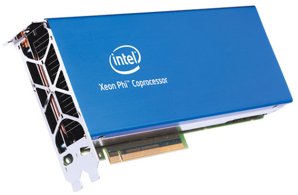
- BEM produces dense matrices
  - Regular, coalesced memory access pattern
- Large BEM matrices approximated using low-rank approximation
  - Large number of dense matrix-vector multiplications
  - Can furthermore benefit from cache hierarchy
- BEM is compute-intensive
  - Numerical/semi-analytical evaluation of singular integrals require large number of flops
  - Benefits from large number of threads and wide SIMD registers of Xeon and Xeon Phi (co)processors

BEM4I library

SIMD vectorization of numerical/semi-analytical quadrature

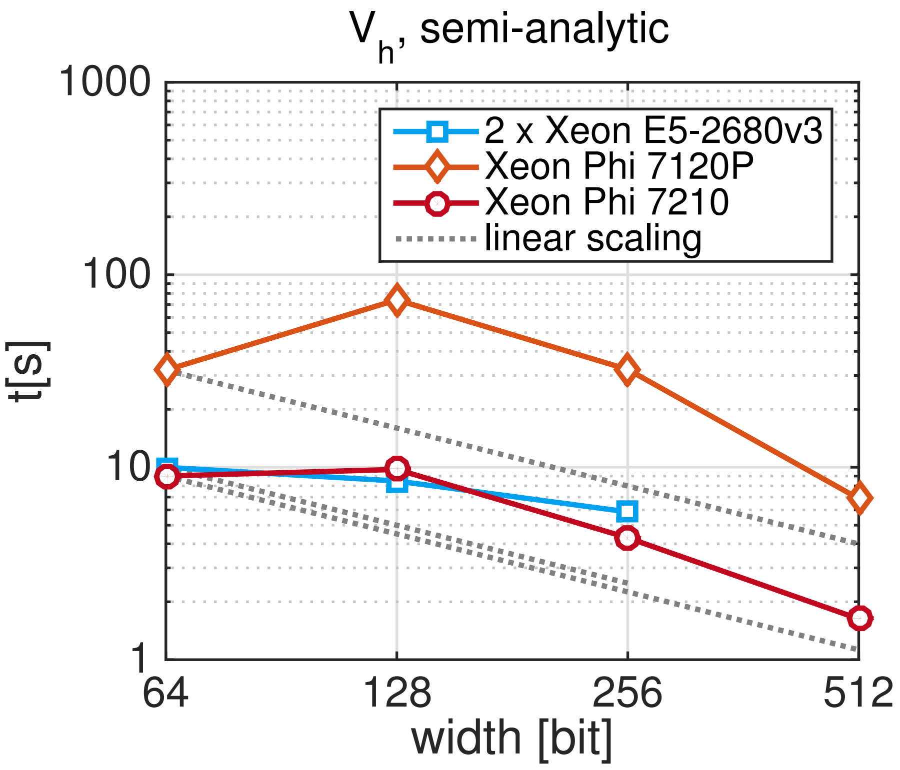
- SIMD instruction sets (AVX512, IMCI) enable concurrent operations on up to 8 DP operands
- Most beneficial optimization techniques
  - OpenMP SIMD pragmas
  - Data alignment and padding
  - AoS to SoA transition for spatial coordinates
  - Unit-strided memory loads and stores
- Tested using
  - 2 x Intel Xeon (Haswell)
  - Intel Xeon Phi 7120P (Knights Corner)
  - Intel Xeon Phi 7210 (Knights Landing)

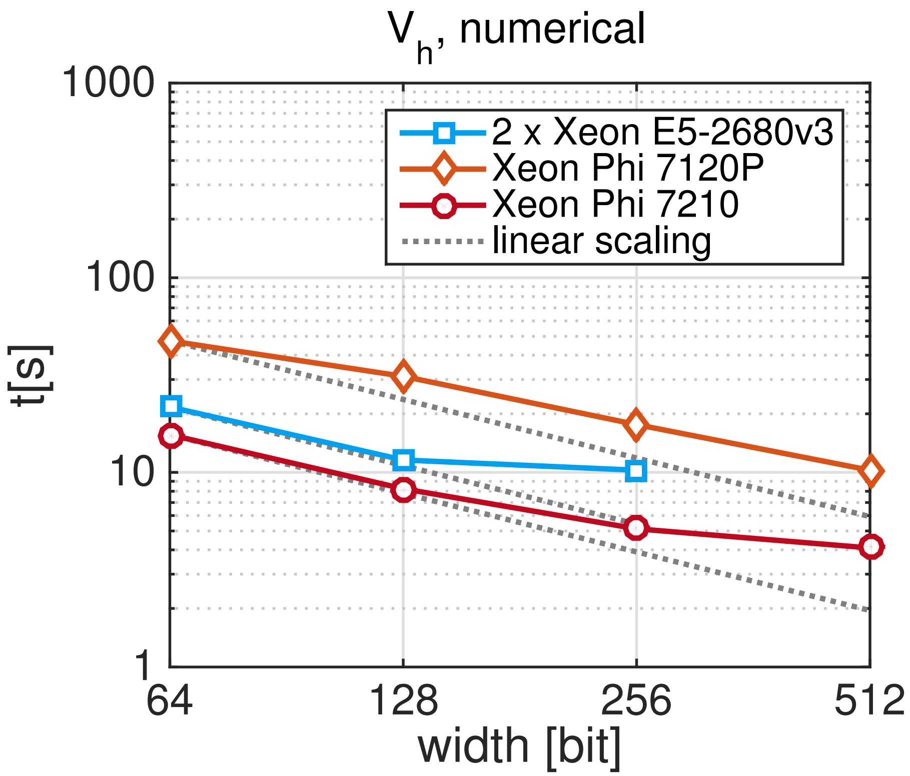
```
#pragma omp declare simd simdlen( 8 )
evaluatePrimitive( double s, ... ) {
...
 // unmasked evaluation of sqrt
 tmp1 = sqrt( tmp1 * tmp1 + q_sq );
 // do not add to f in special case
 if ( abs( s - sx ) > _EPS ) {
  if ( tmp2 < 0.0 ) {
   // masked division only
   tmp3 = hh1 / ( tmp1 - tmp2 );

  } else {

   // masked addition only
   tmp3 = tmp2 + tmp1;

  }


 } else {
  tmp3 = 1.0;
 }
 // unmasked evaluation of log
 f += ( s - sx ) * log( tmp3 );
...}
```



BEM4I library

SIMD vectorization of numerical/semi-analytical quadrature





Performance of AVX2, AVX-512 and IMCI instruction sets on vectors with lengths 1, 2, 4, 8

BEM4I library

SIMD vectorization of numerical/semi-analytical quadrature

	scalar	AVX512(1)	AVX512(2)	AVX512(4)	AVX512(8)
V_h	1.00	1.39	1.29	2.91	7.68
K_h	1.00	1.62	1.72	3.41	8.25

Semi-analytical approach: scalar vs vectorized assembly on Xeon Phi 7210

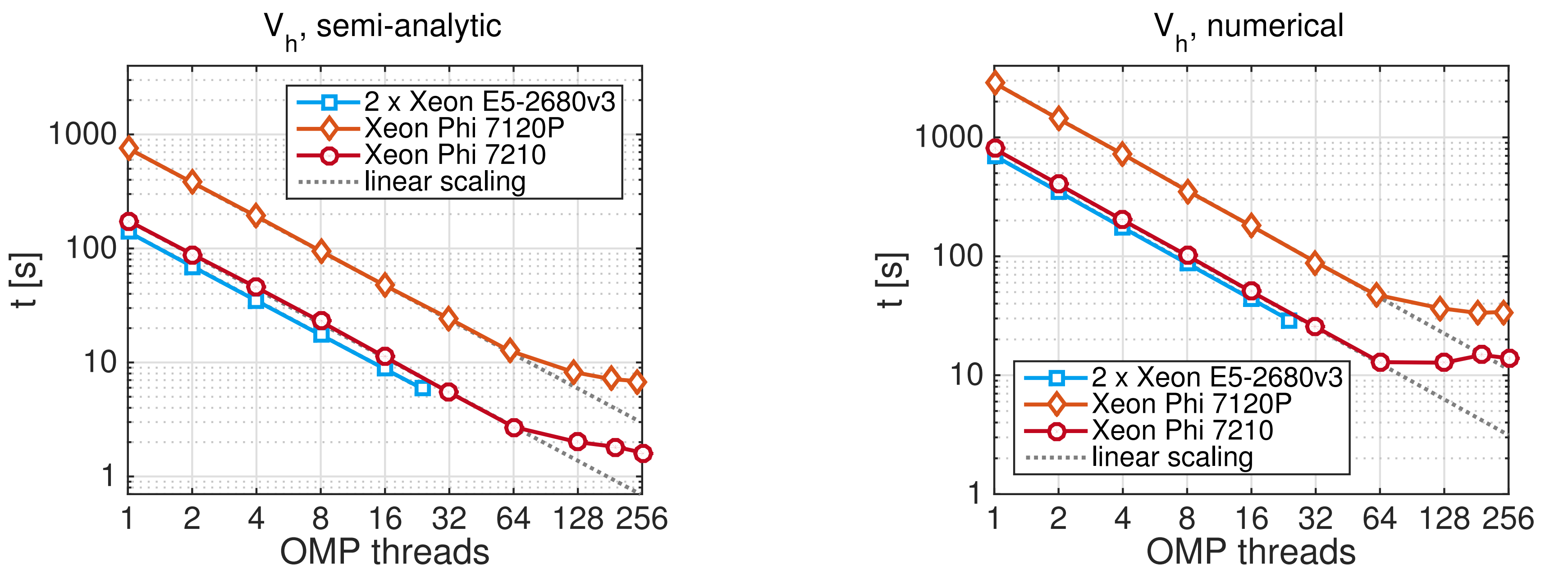
	scalar	AVX512(1)	AVX512(2)	AVX512(4)	AVX512(8)
V_h	1.00	2.00	3.78	6.07	7.62
K_h	1.00	1.20	2.26	3.89	5.53

Fully numerical approach: scalar vs vectorized assembly on Xeon Phi 7210

BEM4I library

Shared memory parallelization

- Using OpenMP parallel pragmas system matrix assembly is distributed among threads
- Tested on 2 x Intel Xeon (Haswell), Intel Xeon Phi 7120P (Knights Corner), Intel Xeon Phi 7210 (Knights Landing)

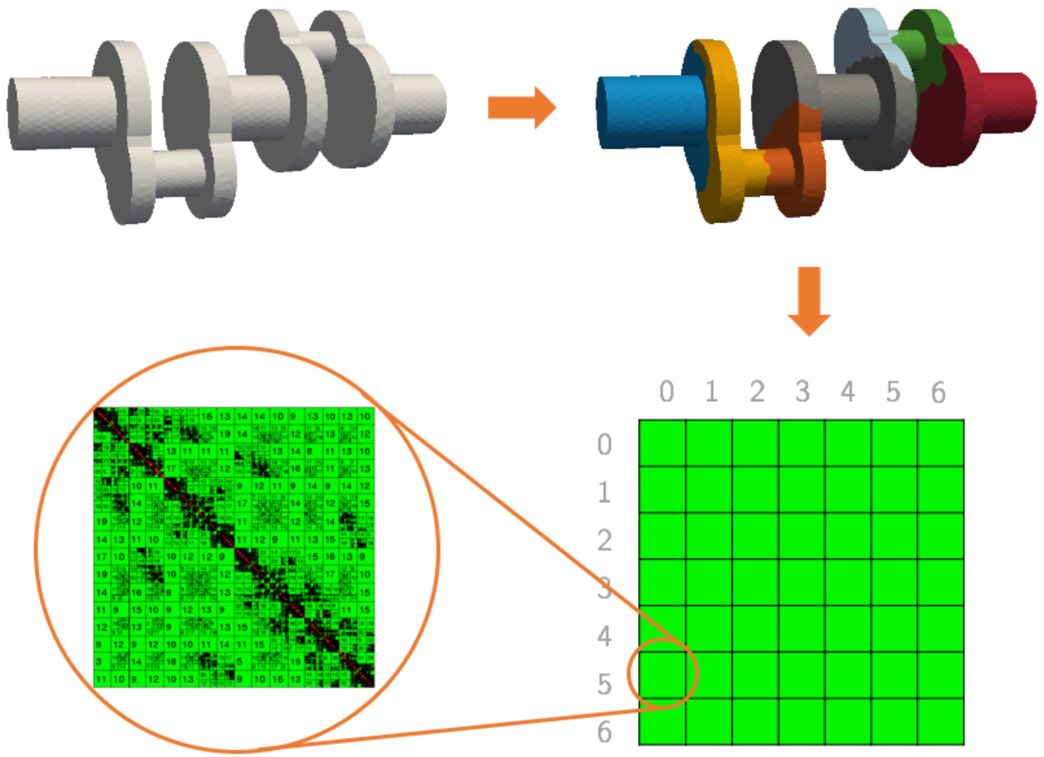


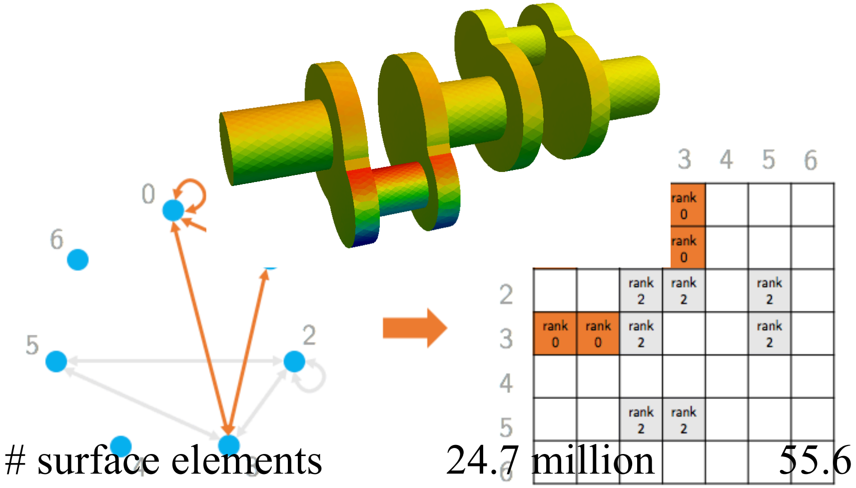
OpenMP scalability on Haswell, Knights Corner and Knights Landing

BEM4I library

Distributed memory parallelization using parallel ACA

- BEM4I implements parallel adaptive cross approximation (ACA)





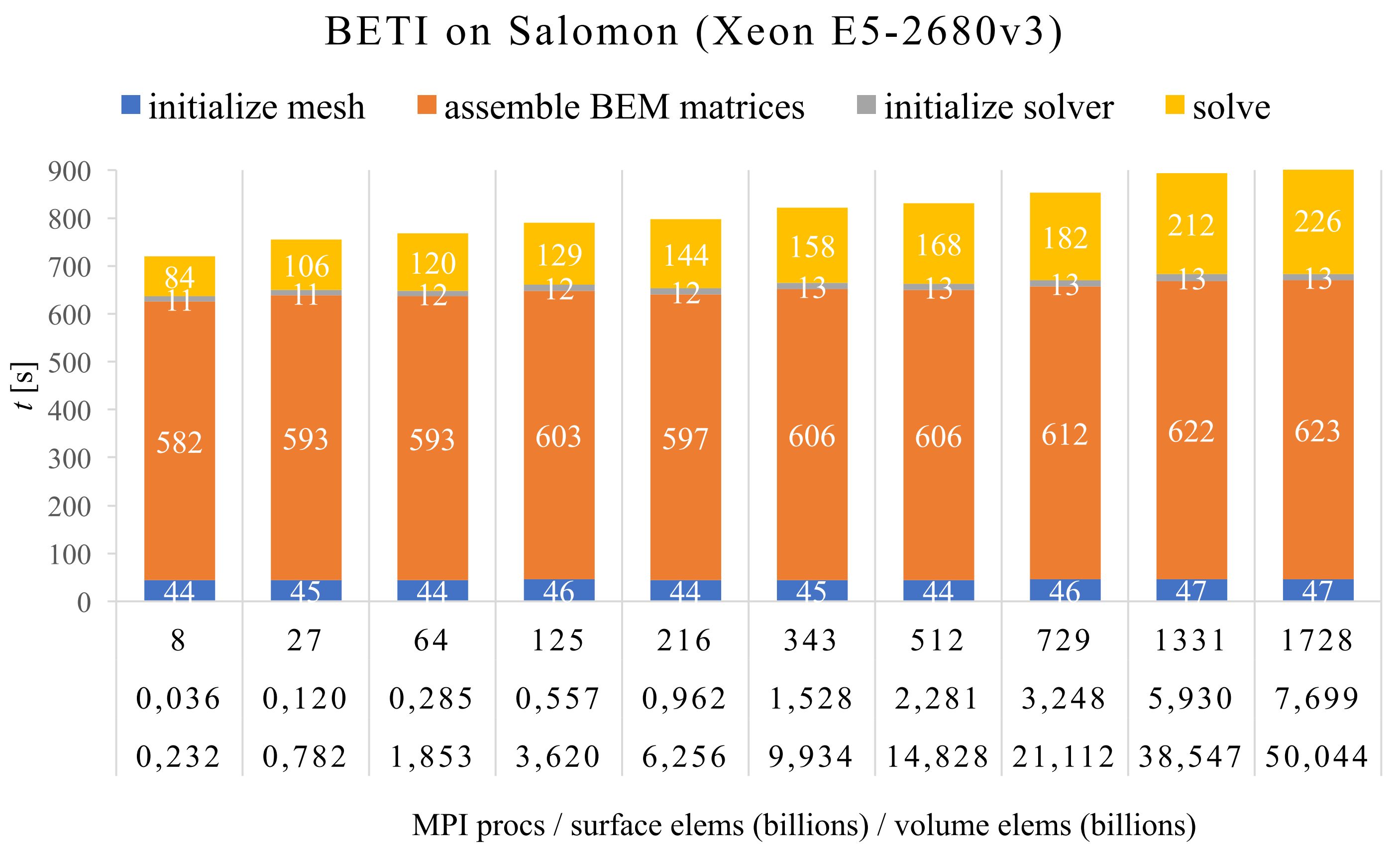
# surface elements	24.7 million	55.6 million
# nodes	64	256
assembly V_h/K_h [s]	163/280	242/414
compr. V_h/K_h [%]	0.08/0.16	0.09/0.18
RAM V_h/K_h [GB]	3932/3987	22688/22873

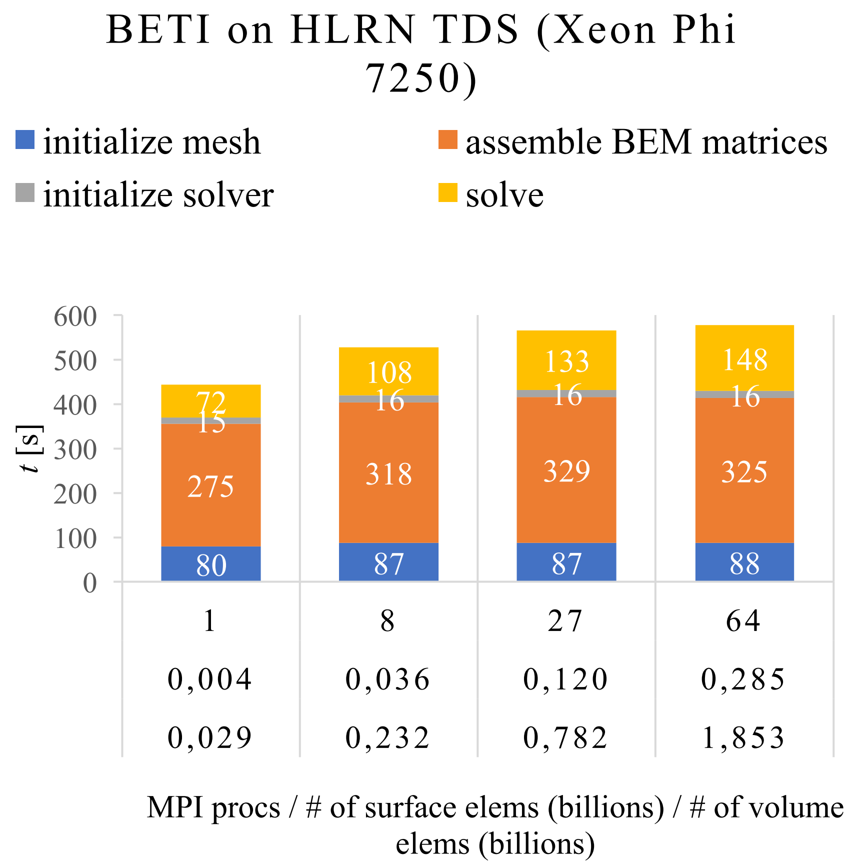
BEM4I library

IT4Innovations
national
supercomputing
center

Distributed memory parallelization using BETI domain decomposition method

- Interface to the ESPRESO library enables us to use boundary element tearing and interconnecting method (BETI)





BEM4I library

Conclusion

- BEM4I provides BEM kernels optimized for Intel architectures
- SIMD and shared-memory parallelization using OpenMP
- Distributed memory parallelization by parallel ACA or BETI DDM
- Suitable for problems on unbounded domains (e.g., sound scattering), shape optimization, etc.
- International cooperation
  - Cooperation with TU Graz (parallel ACA, time-domain BEM, BETI, etc.)
  - Cooperation with Elmer (alternating Dirichlet-Neumann method, proof of concept)
  - Cooperation with UPMC Paris VI (DD based on MTF for the Helmholtz equation)

IT4Innovations national supercomputing center

Agent 327: Operation Barbershop

Cosmos Laundromat: First Cycle

The Daily Dweebs

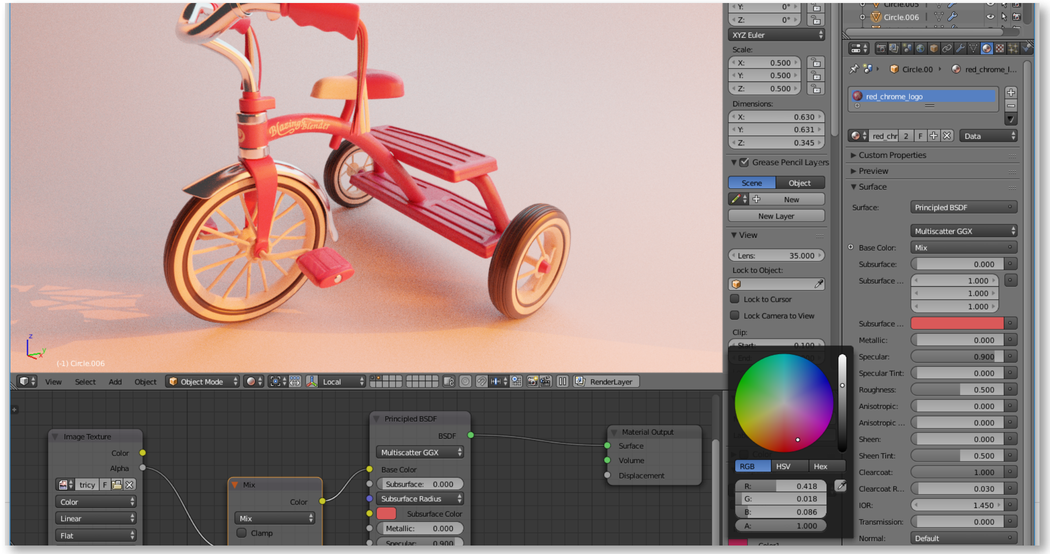
CyclesPhi Render

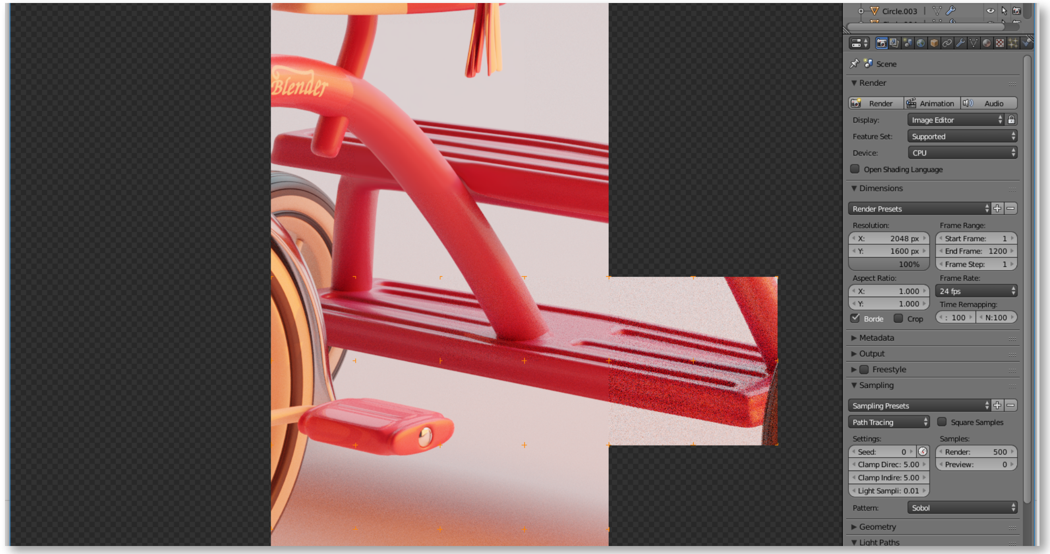
Milan Jaroš

Blender Cycles

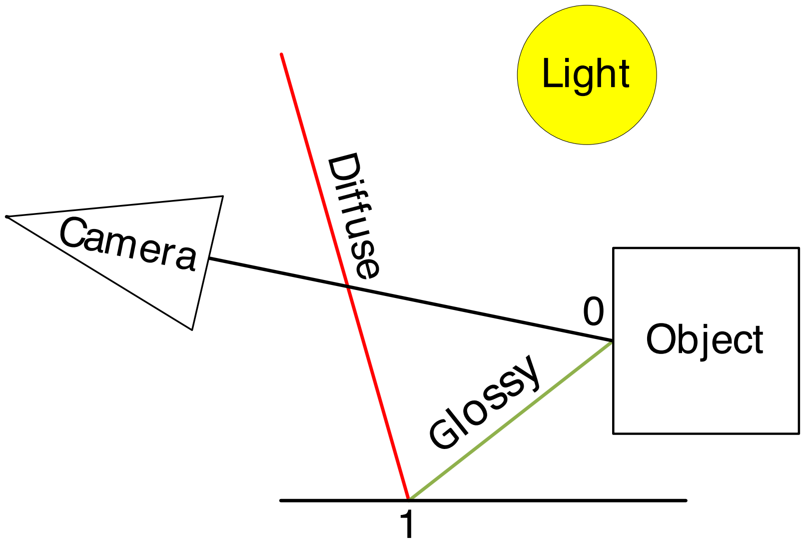
IT4Innovations
national
supercomputing
center

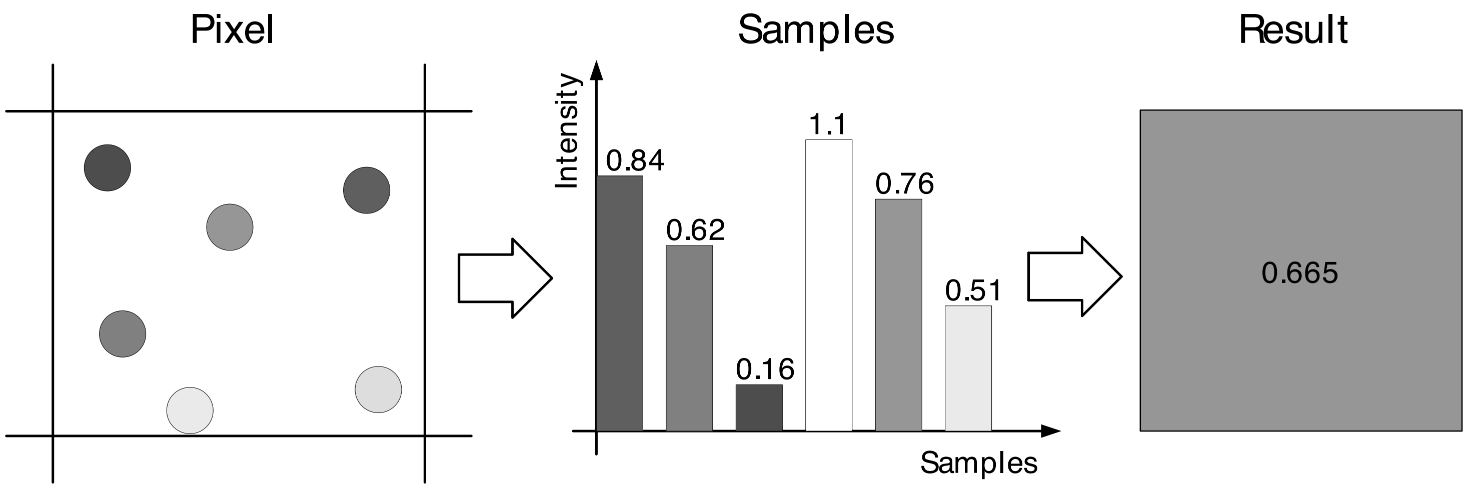
- Blender
  - open source 3D creation
  - render engines: Blender Internal and Cycles.
- Cycles
  - path-tracing based render engine
  - contains shading node system, texture workflow
  - interactive rendering
  - offline renderer





Path Tracing

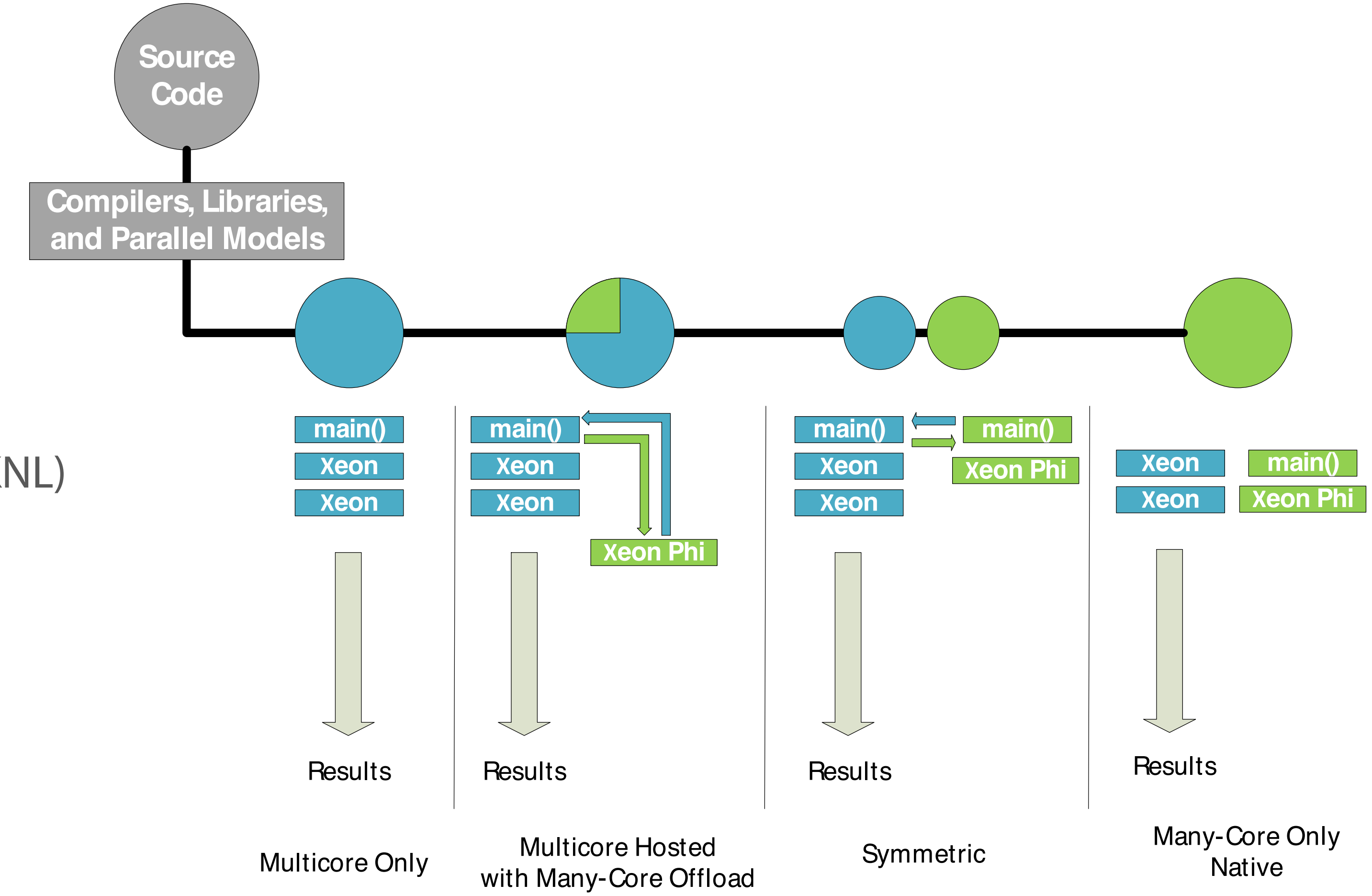




CyclesPhi

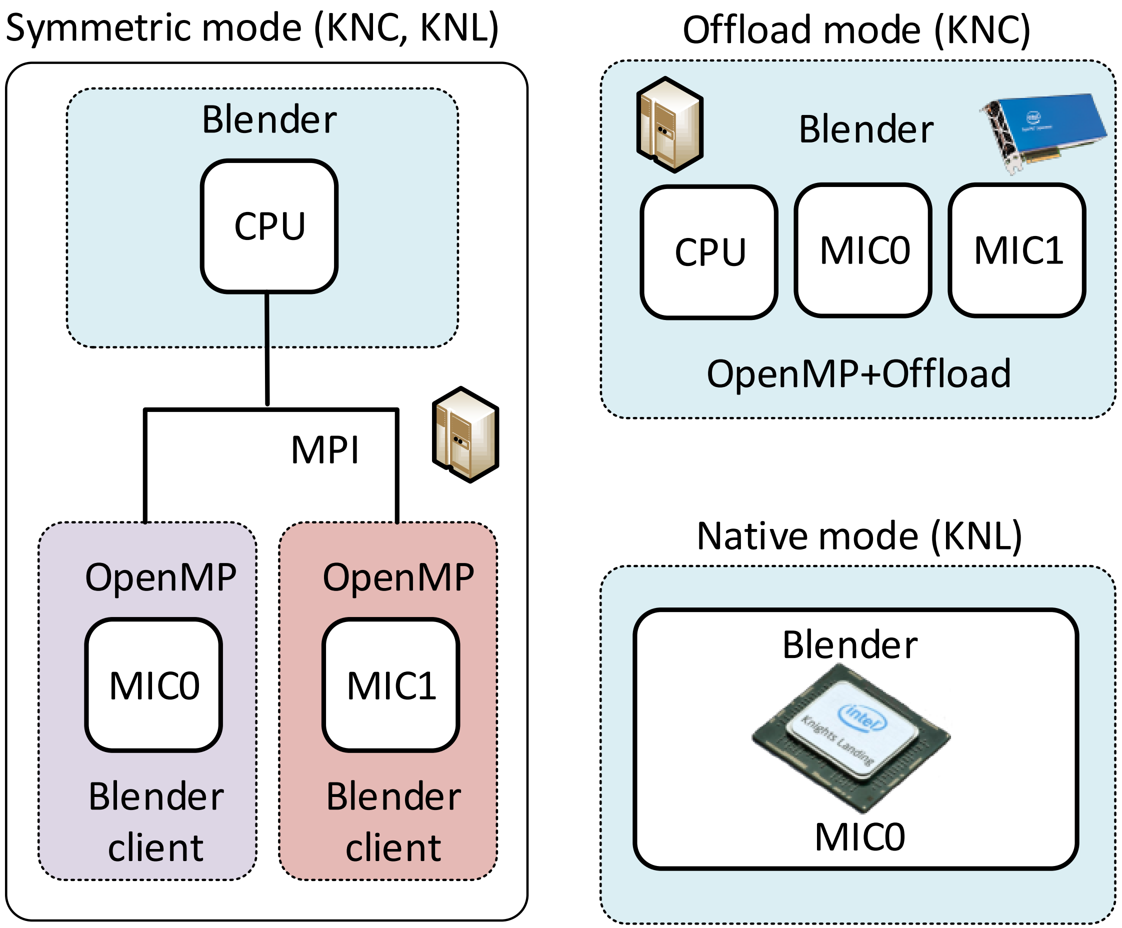
Blender Cycles rendering engine was extended to support HPC environment:

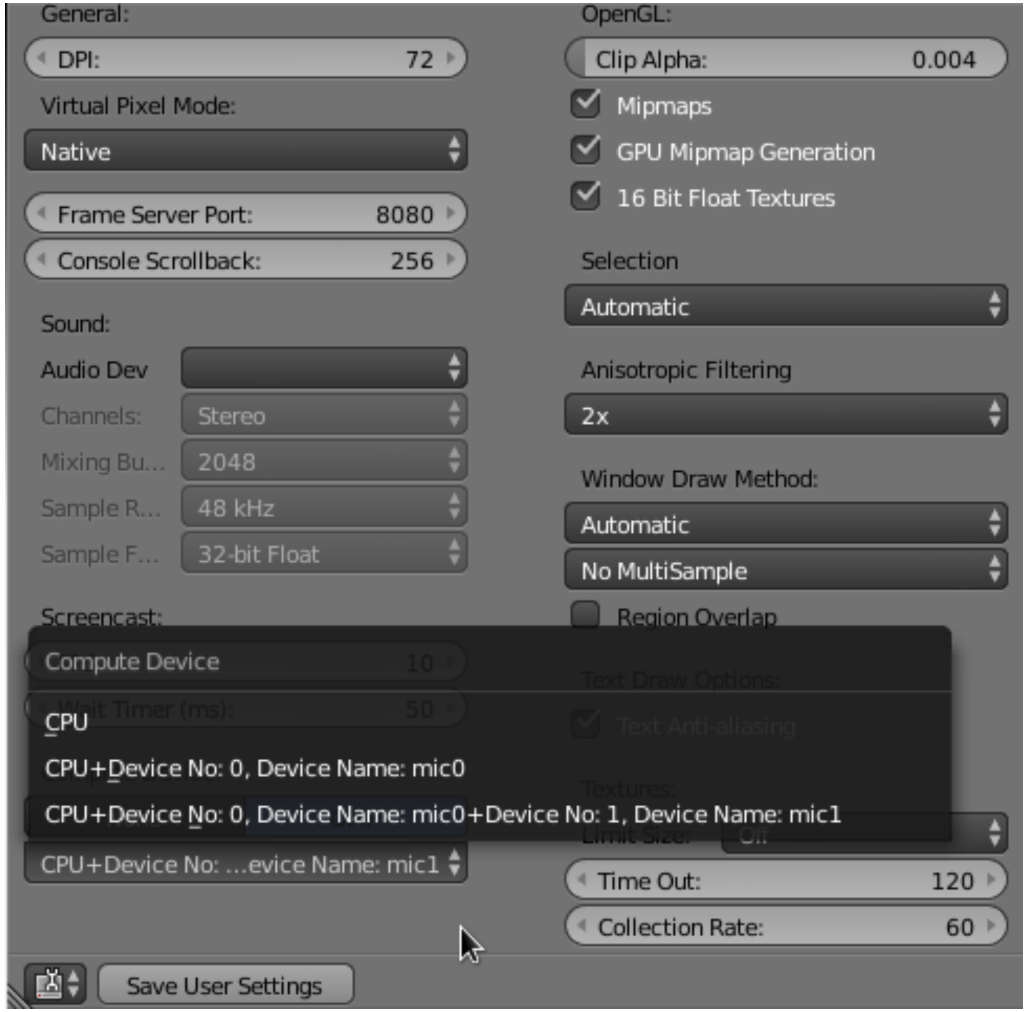
- OpenMP
- MPI
- Intel® Xeon Phi™
  - with Offload concept (KNC)
  - with Symmetric mode (KNC, KNL)
- And their combinations



Native Mode, Offload Mode and Symmetric Mode

IT4Innovations national supercomputing center





Benchmarks

- **Hardware setup**
  - Salomon supercomputer
    - 2x Intel Xeon E5-2680v3 CPUs and 2x Intel Xeon Phi 7120P per node
  - Intel Xeon Phi 7250 at HLRN-III Cray System
  - NVIDIA GeForce GTX 970

Classroom

Classroom by Christophe Seux

Fishy cat

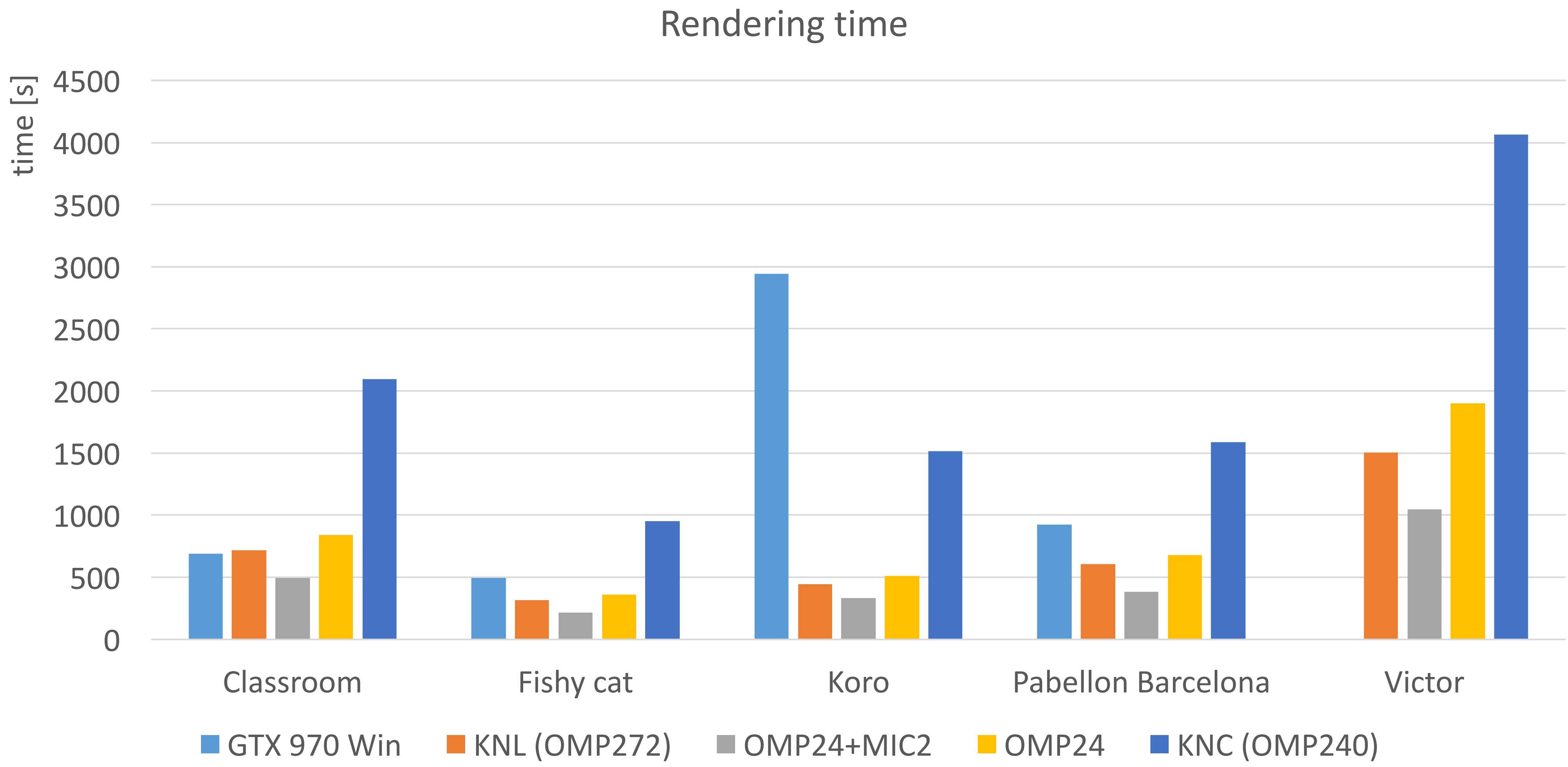
Fishy Cat by Manu Jarvinen

Victor

Victor by Blender Foundation

Performance Comparison

IT4Innovations
national
supercomputing
center

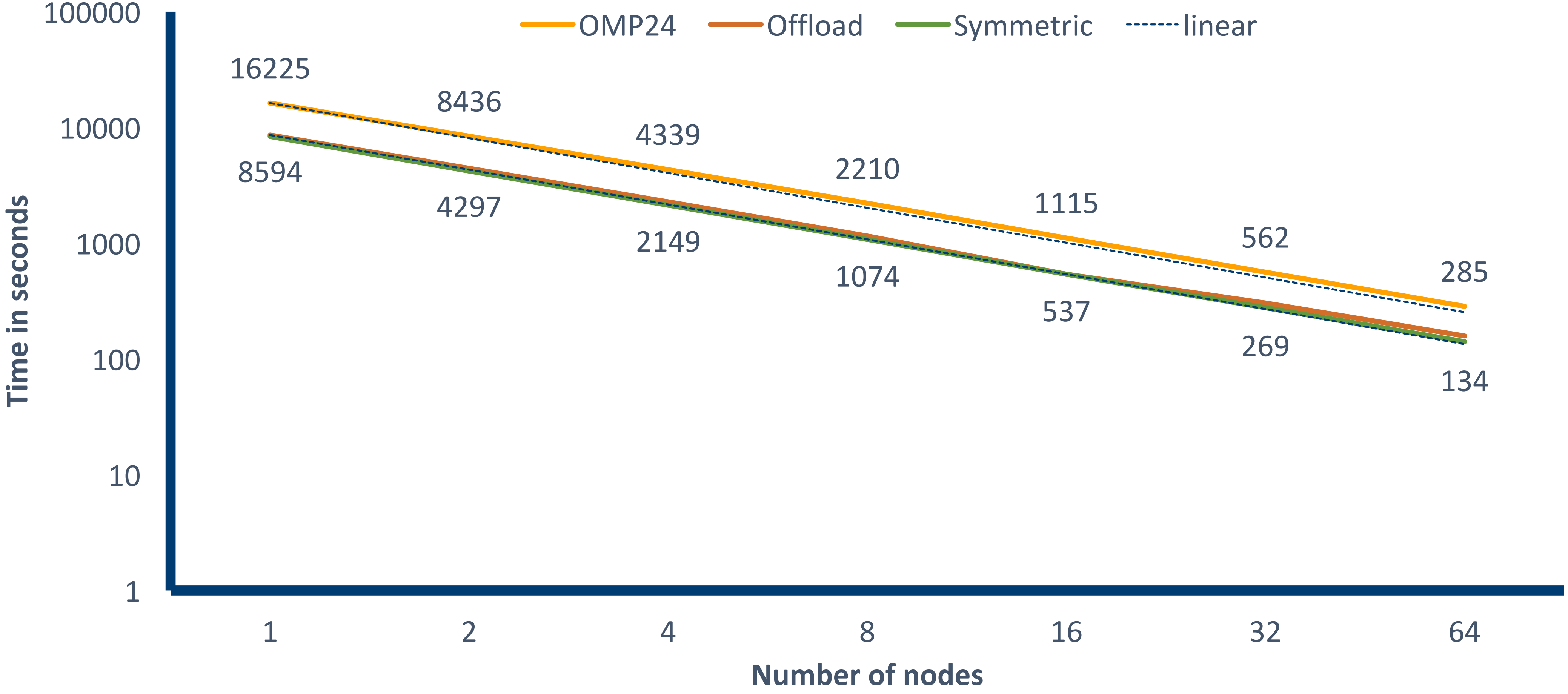


Benchmark Worm: Strong Scalability MPI Test (offline)

- Up to 64 computing nodes of the Salomon supercomputer
- 13.2 million triangles
- Resolution: 4096x2048
- 1024 samples:

Cosmos Laundromat - First Cycle

Benchmark Worm: Strong Scalability MPI Test (offline)



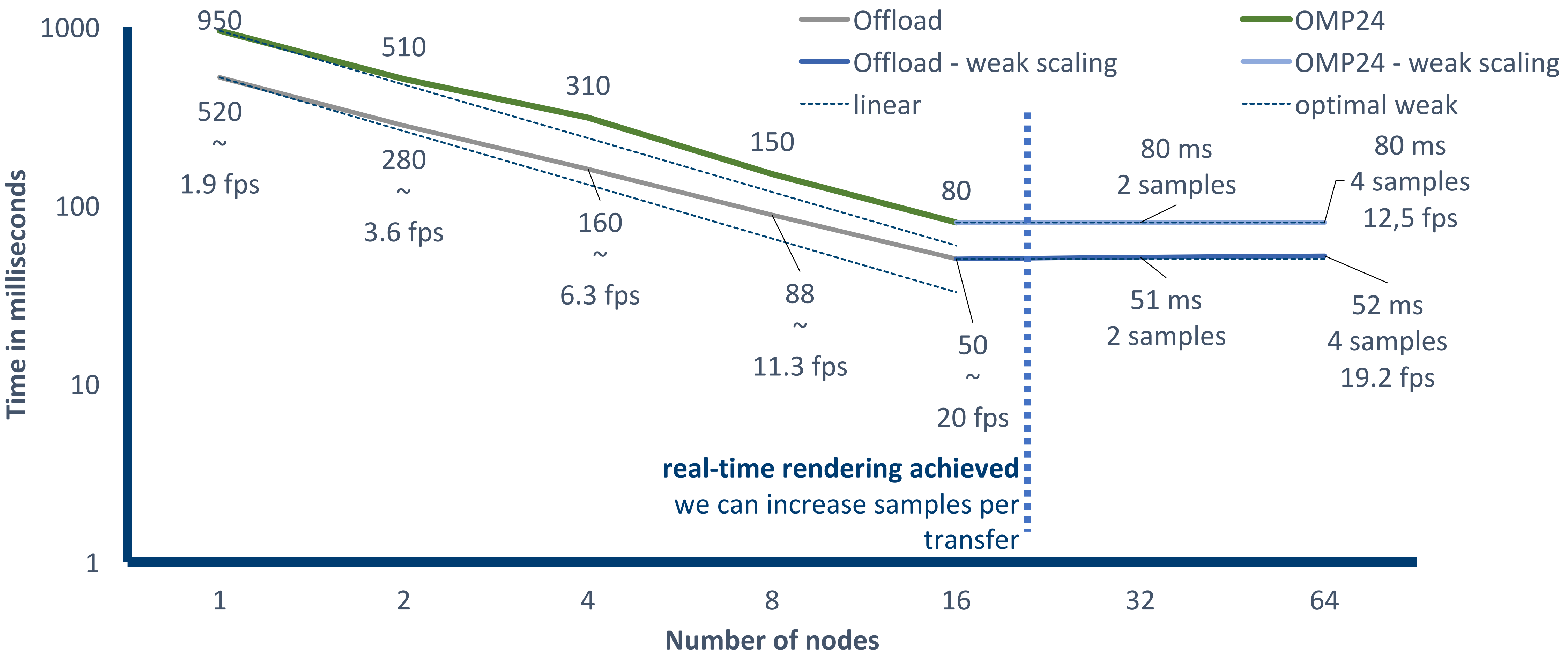
Benchmark Tatra T87: Strong Scalability MPI Test (interactive)

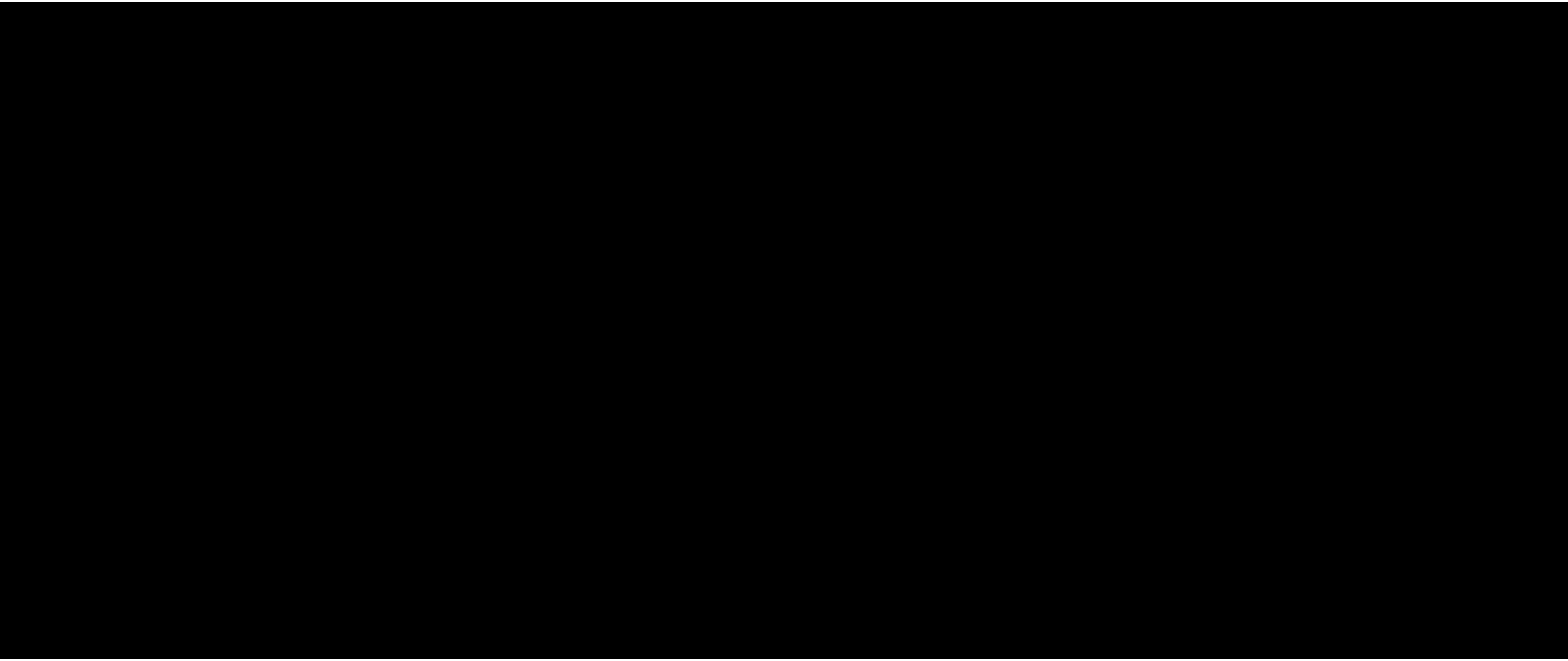
- Up to 64 computing nodes of the Salomon supercomputer
- 1.2 million triangles
- HDRI lighting
- Resolution: 1920x1080,
- 1 sample

Tatra T87 by David Cloete

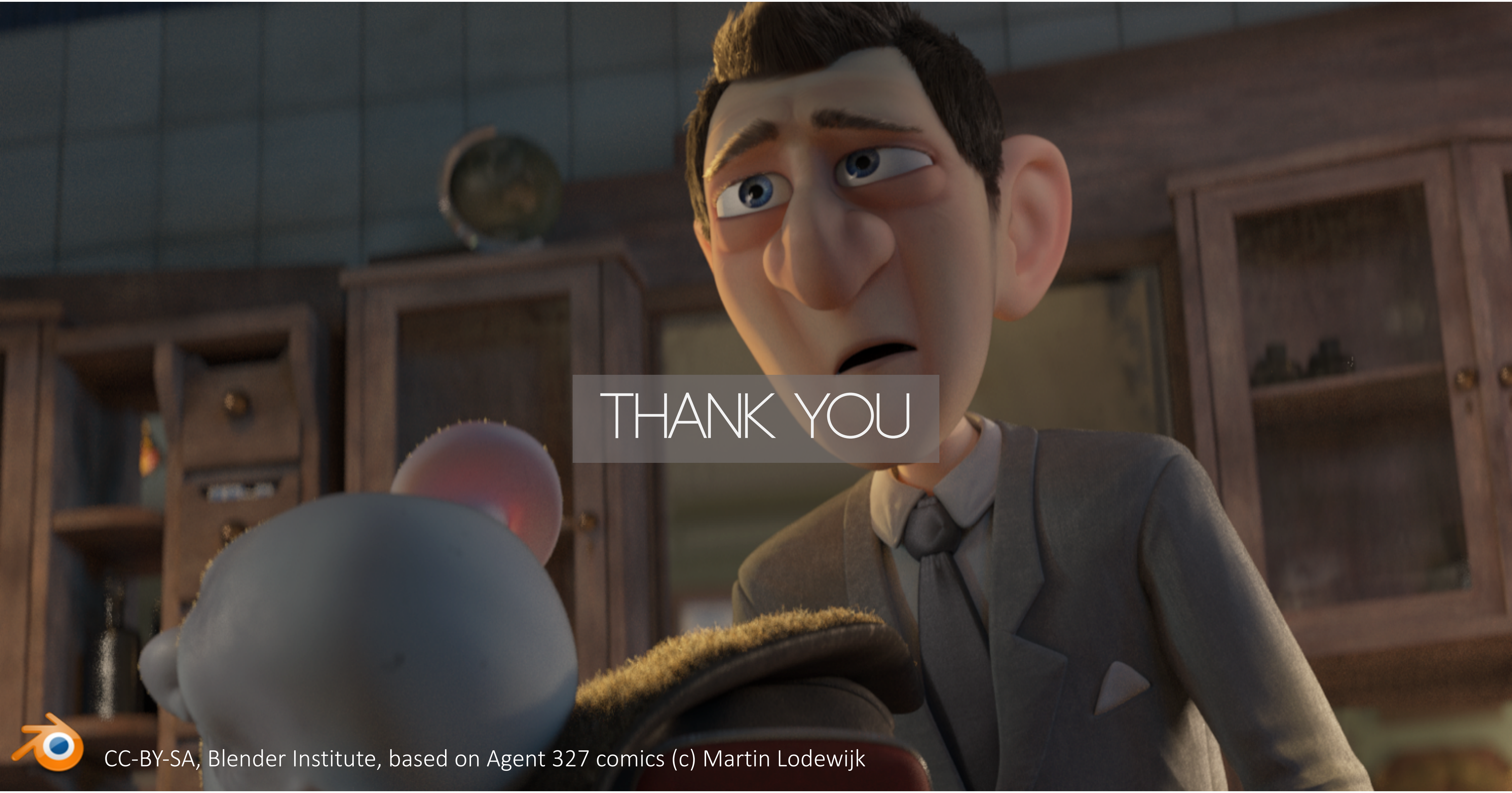
Benchmark Tatra T87: Strong Scalability MPI Test (interactive)

IT4Innovations
national
supercomputing
center





THANK YOU


THANK YOU
CC-BY-SA, Blender Institute, based on Agent 327 comics (c) Martin Lodewijk