Science & Technology Facilities Council

Code_Saturne on Petascale Machines

Charles Moulinec

charles.moulinec@stfc.ac.uk

1st Daresbury-Juelich Workshop 7-8 May 2013

Motivation

- **Mesh generation. Need for parallel mesh generators, but not many are open-source. Do alternative ways exist and are they reliable enough? What about Mesh Joining and Mesh Multiplication, for instance?**
- **Partitioning tools. Due to the large size of the meshes (several billions of cells/elements), partitioning has to be run in parallel. Is there any of the existing tools that is better than the others?**
- **Solving the PDEs. In CFD, iterative solvers are mainly used, but how they scale on high-end machines is still an unknown. Moreover, will MPI alone be enough to get the best of the processors which will be run on?**
- **Input/Output. MPI-IO has been sucessfully used on Terascale machines, but is that the way forward on Petascale ones? Do alternatives exist?**
- **Postprocessing – Restart. The challenge of big data is beginning to be recognised, can some of this stage be moved to the solver for on-the-fly analysis?**

Contents

Science & Technology Facilities Council

Machines Specifications

	HECToR	Jaguar	Blue Joule
Architecture	XE6	XK6	PowerPC
Processor	2 × 16-Core AMD	16-Core AMD	4 quad-core
Nodes	2,816	18,688	7,168
Cores/Nodes	32	16	16
Memory/node	32GiB	32GiB	16GiB
Memory/core	1GiB	2GiB	1GiB
Compiler	GNU 4.6.3	PGI 12.1.0	XL (Fortran V14.1, C V12.0)
MPI	MPICH2 5.4.5	MPICH2 5.4.1	MPICH2 1.4.1p1

TOP500 ranking:

HECToR (Cray XE6): Nov 2009 #20, Nov 2012 #35

Jaguar (Cray XK6): Nov 2009 #1, Not operating anymore

BlueJoule (IBM Blue Gene/Q): Jun 2012 #13, Nov 2012 #16

Science & Technology Facilities Council

Code_Saturne

Technology

- Co-located finite volume, arbitrary unstructured meshes, predictor-corrector method
- 500 000 lines of code, 49% FORTRAN, 41% C, 10% Python
- MPI - OpenMP

Physical modelling

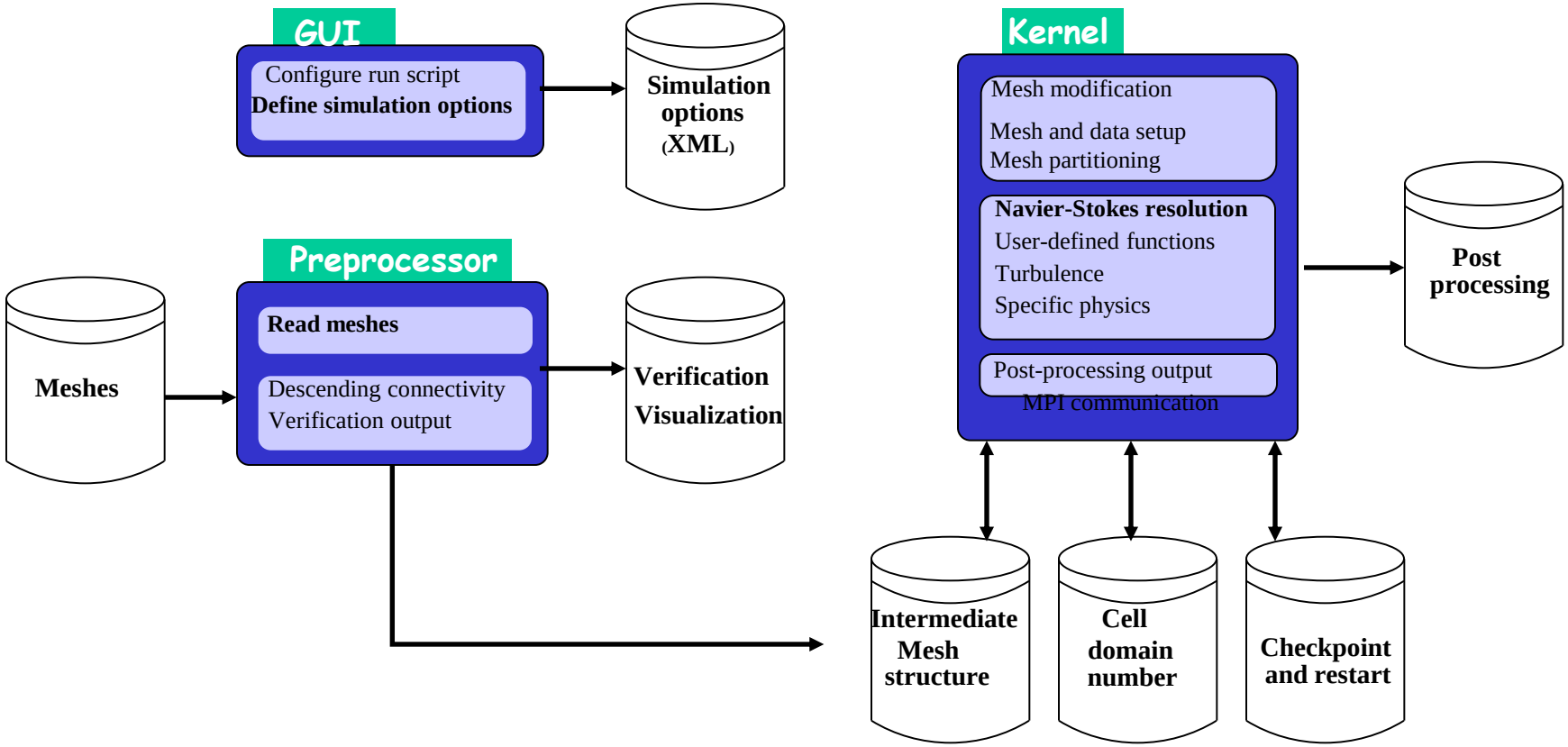
- Single-phase laminar and **turbulent flows**: k-ε, k-ω SST, v2f, RSM, LES
- **Radiative** heat transfer (DOM, P-1)
- **Combustion** coal, heavy fuel oil, gas (EBU, pdf, LWP)
- **Electric** arc and Joule effect
- **Lagrangian** module for dispersed particle tracking
- **Compressible** flow
- **ALE** method for deformable meshes
- **Conjugate heat transfer** (SYRTHES & 1D)
- Specific **engineering modules** for nuclear waste surface storage and cooling towers
- Derived version for **atmospheric flows** (*Mercure_Saturne*)
- Derived version for **eulerian** multiphase flows

Flexibility

- **OpenSource**
- **Portability** (UNIX, Linux and MAC OS)
- **GUI** (Python TkTix, Xml format)
- **Parallel** on distributed memory machines
- **Periodic boundaries** (parallel, arbitrary interfaces)
- Wide range of **unstructured meshes** with **arbitrary interfaces**
- **Code coupling** capabilities (*Code_Saturne*/*Code_Saturne*, *Code_Saturne*/*Code_Aster*, ...)

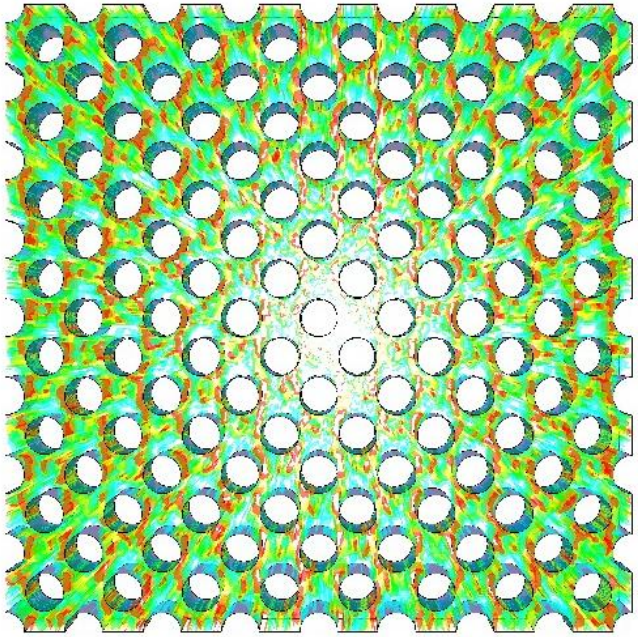
Science & Technology Facilities Council

Code_Saturne



Test Case

Large-Eddy Simulations in staggered-distributed tube bundles.



Experiment of Simonin and Barcouda.

2-D section: 100,040 cells; 3rd direction: 128 layers -> 13M cells

Science & Technology Facilities Council

Mesh Joining

Split the computational domain in N parts and mesh each part independently. Joining might be non-conforming.

Time to join 4 x 812M hexa- cell meshes

conforming

Cores	HECToR	Jaguar
8192	95	127
16384	146	178
32768	400	414
65536	1560	1630

Time to join 15 x 108M tetra - cell meshes: 23 s (HECToR Phase2b)

3072 MPI tasks using 4GiB RAM each.

non-conforming

Science & Technology Facilities Council

Mesh Multiplication

From a coarse grid, split the cells/elements homogeneously

Special treatment is required to preserve the surface description

Time to generate a 26B cell mesh from a 51M cell mesh, for the tube bundle case (hexahedral cells only) (3 levels)

Cores	Blue Joule
16,384	23.7s
32,768	14.5s

Time to generate a 105B cell mesh from a 26M cell mesh, for the tube bundle case (hexahedral cells only) (4 levels)

Cores	Mira
262,144	8.46s
524,288	8.95s

Argonne NATIONAL LABORATORY

Graph Partitioning

The most popular parallel distributions are ParMETIS and PT-SCOTCH.

Taking the example of ParMETIS, it is based on a multilevel k-way partition algorithm that works in three steps, with first the coarsening which involves a random process, then the partitioning of the coarse graph by a k-way partition and finally the uncoarsening.

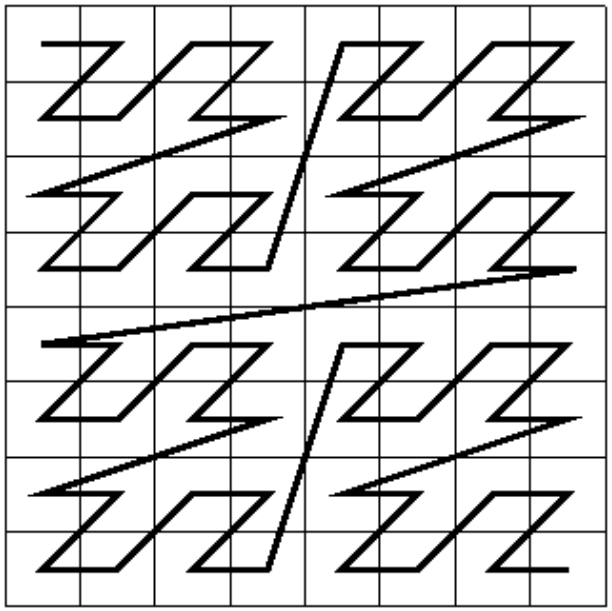
Both tools might work on a number of processors which is smaller than the number of sub-domains required for the simulation, but for a given number of sub-domains, the quality of the partition is in general better when the number of processors ParMETIS is run on is small.

Reducing their number of processors means that the total memory required is distributed between less processors, which is a clear limitation with only 1GiB RAM per core. A compromise has to be found, and the number of processors to run graph-based tools is usually equal to the number of subdomains.

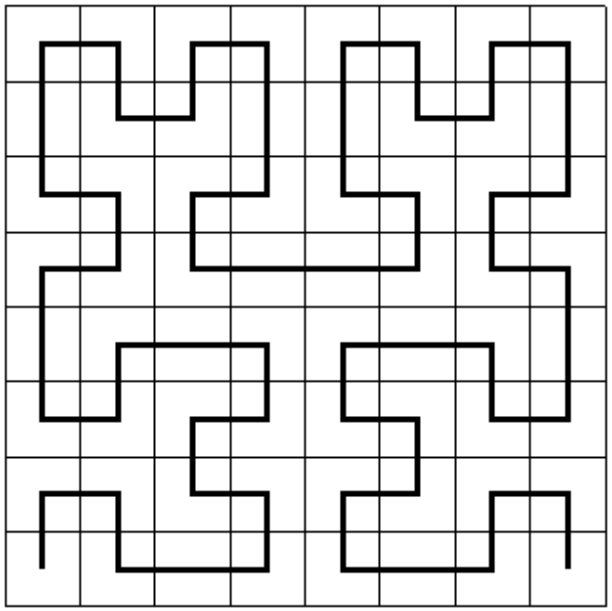
Geometry Partitioning

Principle: Mapping between 2-D/3-D and 1-D

Based on Space Filling Curves



SFC Morton



SFC Hilbert

Partitioning-Results

Test case: 3.2B cell mesh

Cores	Partitioning				Halos			
	HECToR		Jaguar		HECToR		Jaguar	
	SFC	ParM	SFC	ParM	SFC	ParM	SFC	ParM
8192	21.7	45.6	14.8	14.3	10.6	8.2	11.1	8.7
16384	28.2	90.7	9.7	16.1	9.5	8.3	8.9	7.8
32768	53.5	181.0	16.4	40.8	22.6	21.2	18.7	17.7
65536	97	-	57.6	94.1	94.9	-	81.3	63.1

For 65536 cores, ParMETIS needs >1GiB, impossible on HECToR.

-SFC Morton usually faster.

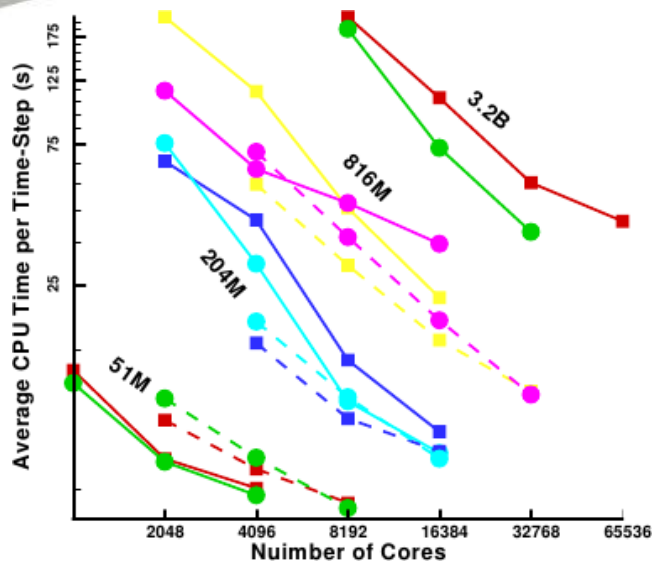
Computing the halos requires more time when SFC Morton is a partitioning tool, probably due to the poorer edge-cut quality.

Science & Technology Facilities Council

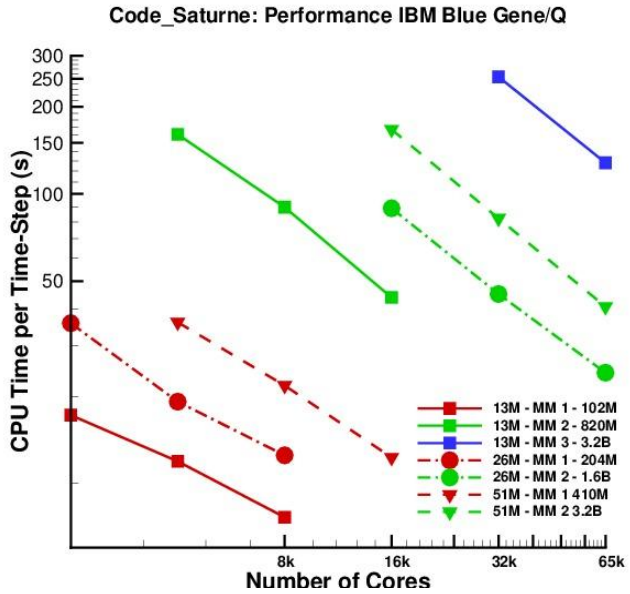
Solving the PDEs

HECToR – Blue Joule

Mesh Joining

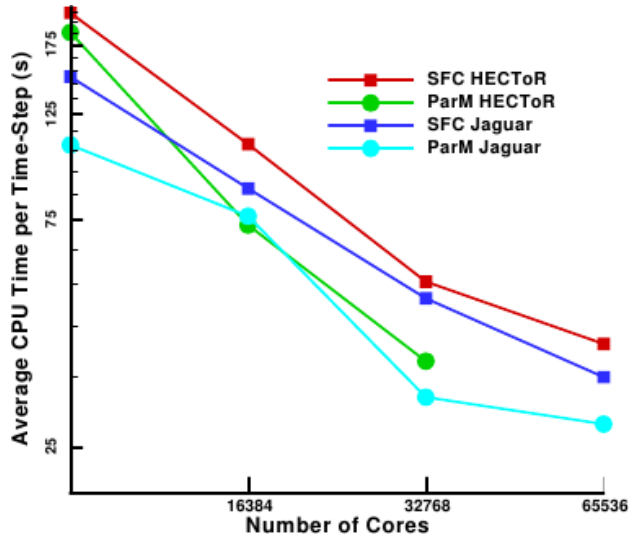


Mesh Multiplication



HECToR - Jaguar

Mesh Joining



Cores	Mira
262,144	1,116s
524,288	739s

105B Cell Mesh

Argonne NATIONAL LABORATORY

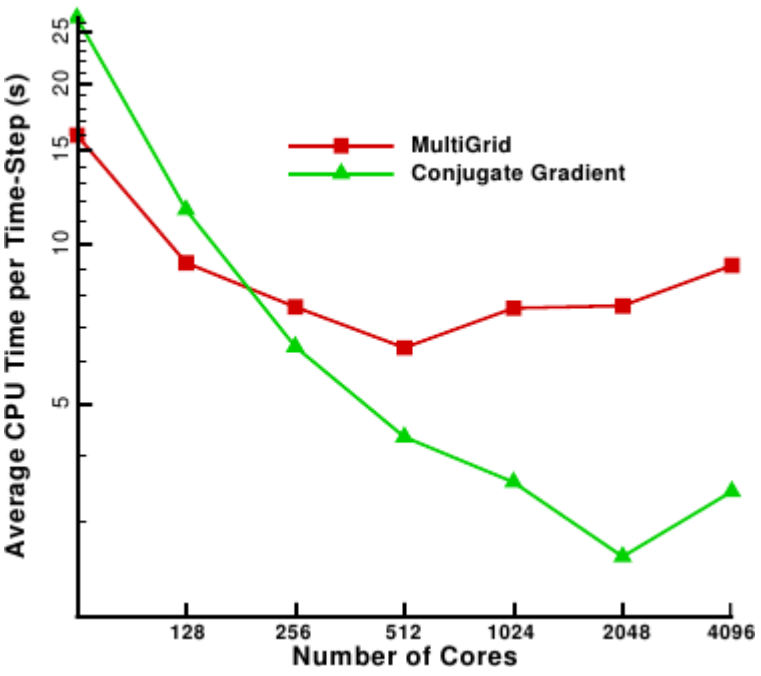
Solving the PDEs

Smaller case (5.7M cells) to compare the behaviour of the multigrid and of the conjugate gradient as a function of the number of processors, when both are used to solve the pressure Poisson equation.

From 256 processors, the conjugate gradient algorithm is faster than the multigrid algorithm. A speed-up is observed up to 2048 cores, whereas no gain in time is observed after 512 cores for the multigrid algorithm.

Moreover, multigrid for something else than elliptic problems is still an open problem.

Probable way forward: multigrid as a preconditioner and deflated conjugate gradient-like method as a solver.



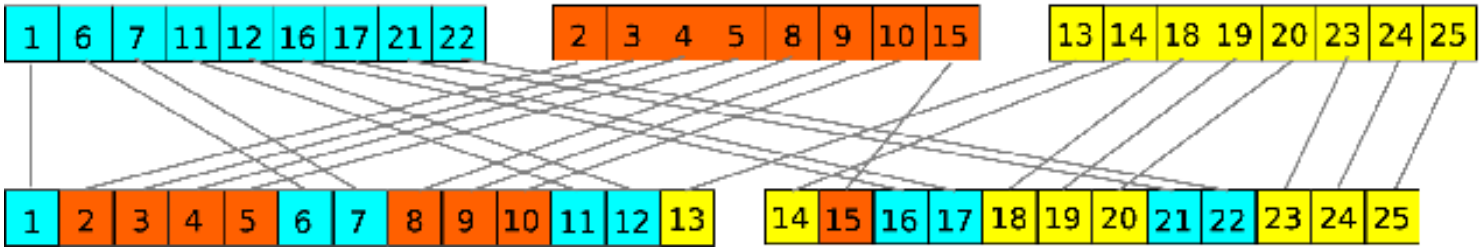
IO Management

Use of global numbering

-Redistribution on n blocks

- n blocks ≤ n cores
- Minimum block size may be set to avoid many small blocks (for some communication or usage schemes), or to force 1 block (for I/O with non-parallel libraries)

21	22	23	24	25
16	17	18	19	20
11	12	13	14	15
6	7	8	9	10
1	2	3	4	5



IO Management

Comparison IO per Blocks (Ser-IO) and MPI-IO

Comparison Lustre (Cray) / GPFS (IBM BlueGene/Q) filesystems

	HECToR		Blue Joule	
Cores	MPI-IO	Ser-IO	MPI-IO	Ser-IO
2048	633	1203	-	-
4096	608	640	85	1279
8192	859	1147	86	1300
16384	732	747	67	1330
32768	-	-	59	1360

Tube Bundle

812M cells

Ser-IO: ~same performance on Lustre and GPFS

MPI-IO: 8 to 10 times faster with GPFS

MPI-IO: about 35 minutes to write a 26B cell mesh file (6TB)

Science & Technology Facilities Council

Postprocessing-Restart

Very big data are generated (>several TB. For instance, an instance of the flow field for the 26B cell case would require about 15TB storage).

Major part of the postprocessing should be performed in parallel by the code.

Two types of data:-

-monitoring points or probes, which are used to assess the quality of the simulation.

Dumping data at probe locations is straightforward, no matter the size of the meshes.

-films. Two strategies are identifed, i.e. :

***generating frames on a much coarser grid than the one used for the resolution of the PDEs and dumping each frame on the disk before combining them there. Natural approach if grid obtained by Mesh Multiplication.**

***linking the solver to a visualisation library. This allows real time visualisation in the course of the simulation, and is of great help for debugging.**

The restarting procedure is the current main bottleneck as all the variables that have to be saved do belong to the finest grid used for the PDEs. It relies on how fast the IO operations are handled.

Final Remarks

- **Mesh generation. If no parallel mesh generators available, Mesh Joining combined with Mesh Multiplication are a good option.**

- **Partitioning tools. ParMETIS in general faster than SFC Morton partitioner.**

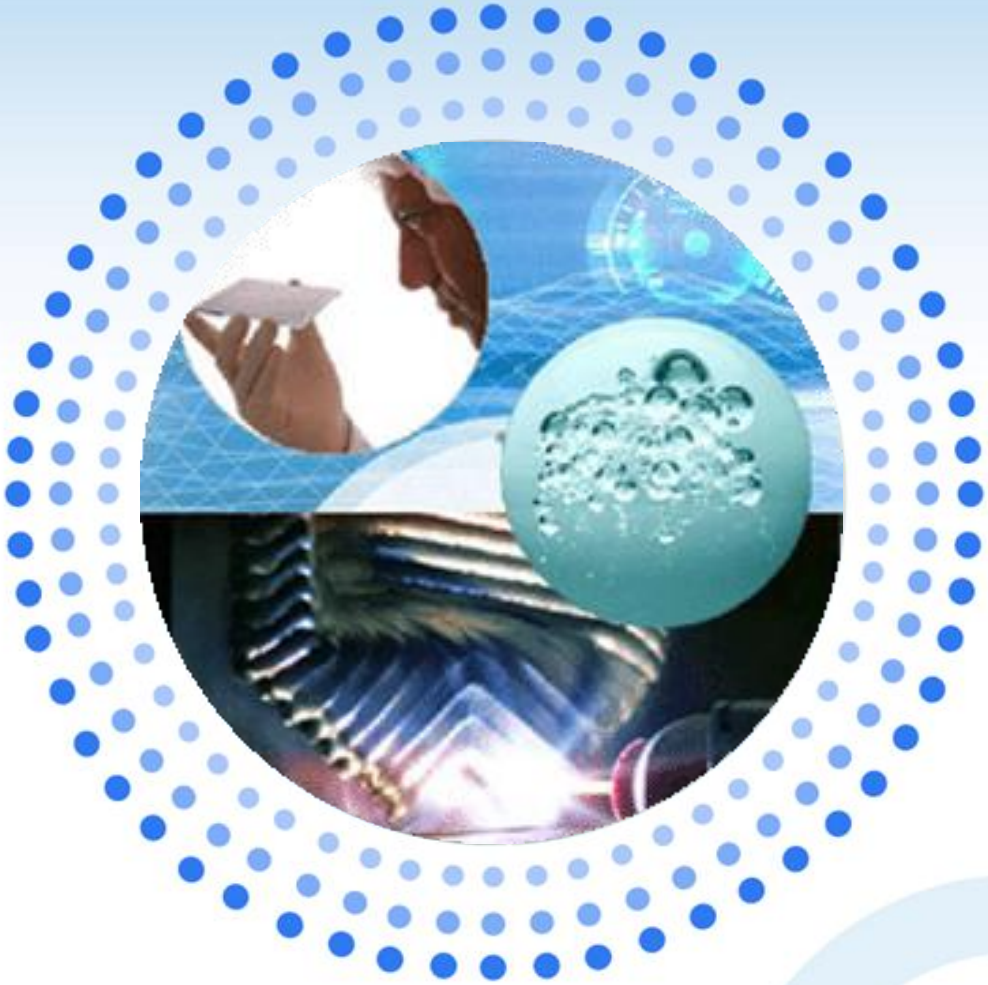
- **Solving the PDEs. Multigrid used for preconditioning and a deflated conjugate gradient for solving.**

- **Input/Output. MPI-IO works well for large dataset, especially on GPFS file system.**

- **Postprocessing – Restart. Monitoring probes and films generated from the solver. Restarting strategy highly depends on the performance of MPI-IO.**

Thank you very much.

The Telemac system



An integrated tool for environmental CFD

Used by EDF R&D for dimensioning and safety studies

Many other users all over the world

EDF R&D

The TELEMAC hydroinformatic system

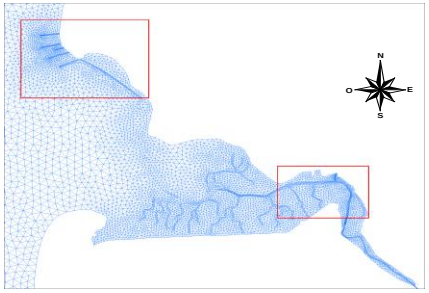
	2 dimensions	3 dimensions	SPH (2D or 3D)
Hydrodynamics	TELEMAC-2D	TELEMAC-3D	SPARTACUS
Sediment	SISYPHE*	TELEMAC-3D	
Water quality	Chaining with Delwaq	Chaining with Delwaq	
Groundwater flows	ESTEL-2D	ESTEL-3D	
waves	ARTEMIS TOMAWAC		

BIEF Finite Elements Library	Fudaa / Rubens / Blue Kenue Tecplot Pre- and post-processors	Matisse / Janet Mesh generators

* Co-property **LNHE-Cetmef**

The Telemac system

Main characteristics



- Developed since 1987 at EDF R&D / LNHE
- World distributed (first commercial with 200 licences, now freeware and open source)
- FORTRAN 90, PERL, MPI
- Based on unstructured grids
- Documentation and validation

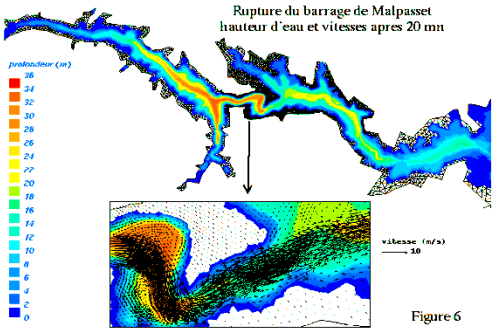
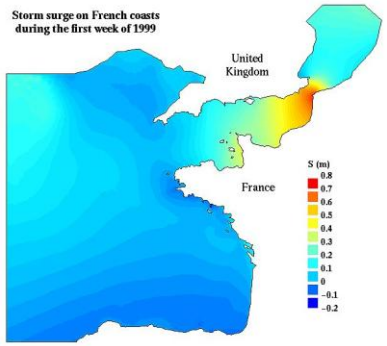
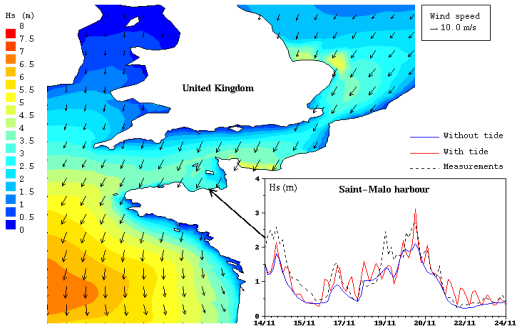


Figure 6



Key features

- Finite Elements, Implicit schemes
- Parallelism with domain decomposition
- Dry zones
- Non hydrostatic 3D with free surface



A co-development strategy

A consortium of core members:

Electricité de France

Cetmef (Ministry of Equipment, France)

BundesAnstallt für WasserBau (Germany)

Sogreah (France)

Hydraulic Research Wallingford (UK)

Daresbury Laboratory (UK)

And many others:

Deltares (ex Delft Hydraulics, Holland)

HydroQuebec (Canada), Centre Hydraulique Canadien, S.H.O.M. (French Navy)

IMFT, INRIA

Hannover, Stuttgart, Karlsruhe,… Universities (Germany)

Manchester, Bristol, Bangor,… Universities (Great-Britain)

http://www.opentelemac.org

3500 registered users

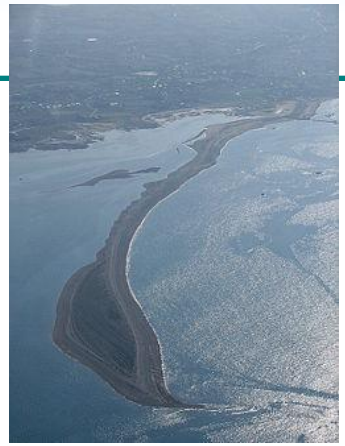
7500 messages on the forum

108 countries

130 persons in 2011 user club in Paris

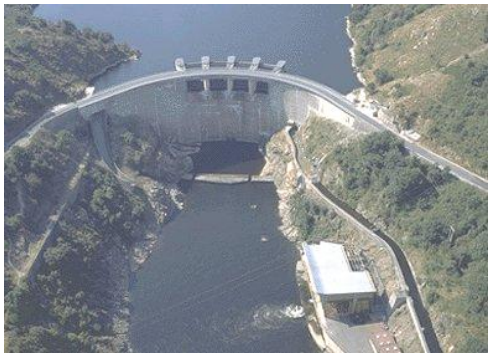
20th user club 16-18 October 2013 in Karlsruhe

The Telemac system



Telemac-2D

- Shallow water equations (Saint-Venant)
- Boussinesq equations
- Meshes of triangles
- Dry zones, turbulence models
- Tracers (temperature, pollutants, etc.)



Tidal prediction and storm surge simulation

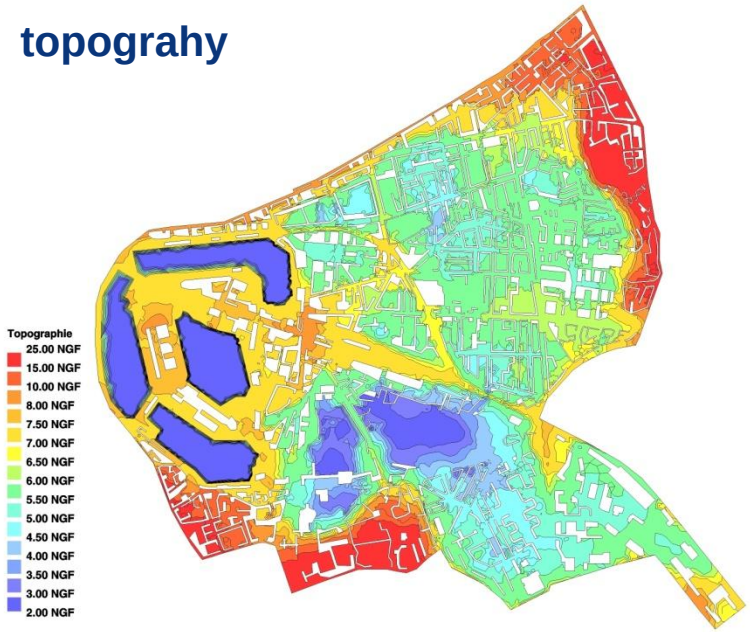
Main objectives

- Navigation safety
- Harbour design and coastal defence
- Flooding in estuaries and coasts
- Designing marine current turbines
- Pollutant advection
- etc.

Flooding in Saint-Malo?

© SOGREAH

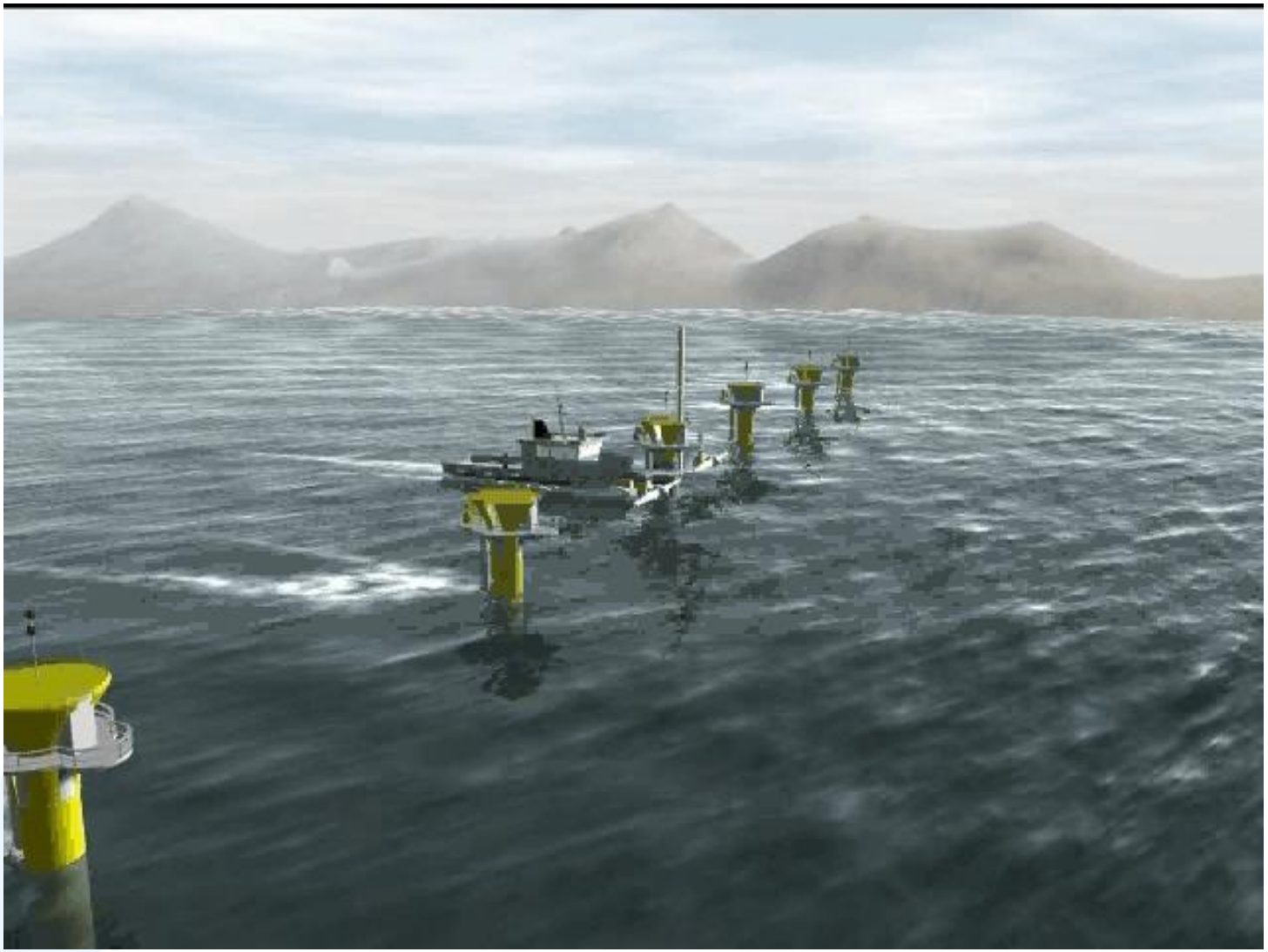
topograhy



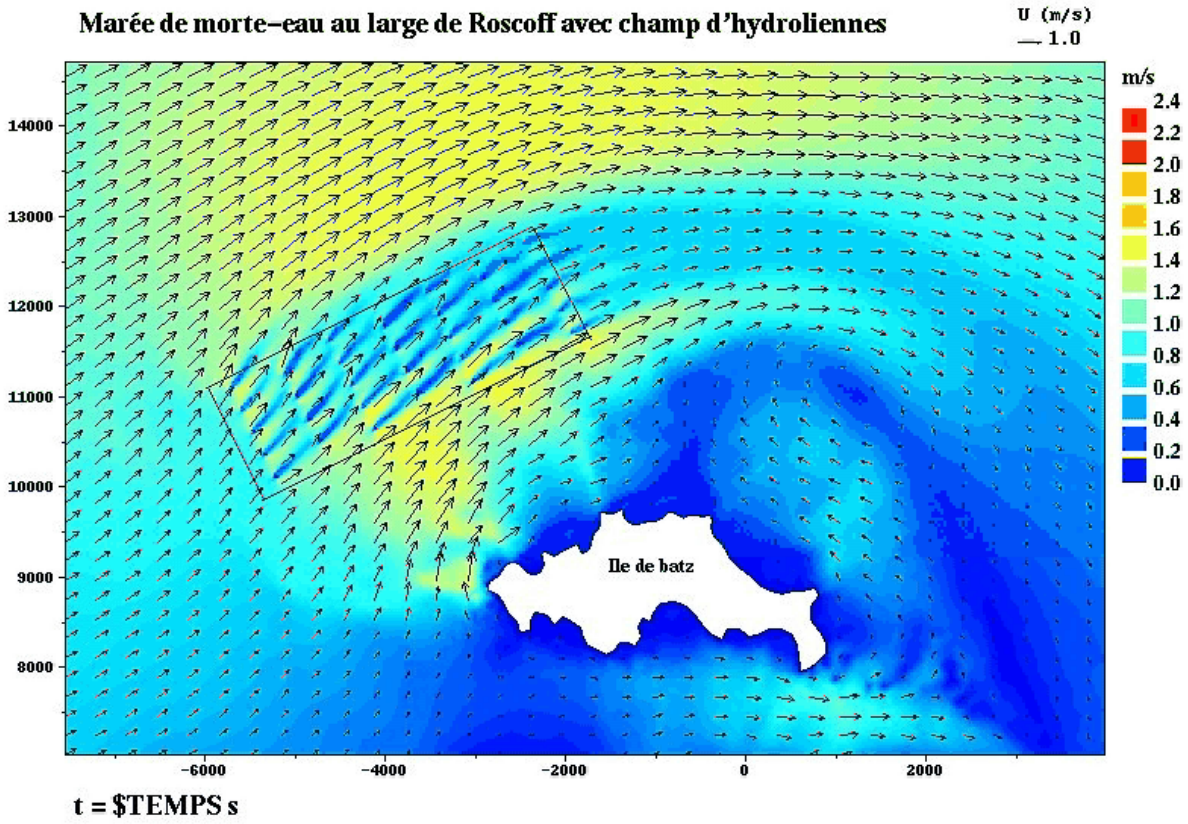
Context

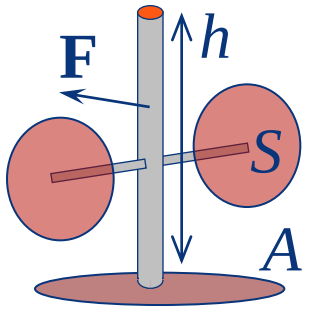
The city of Saint-Malo is subject to severe storm surges combined with high wind waves, capable of strong overtopping through harbour defences

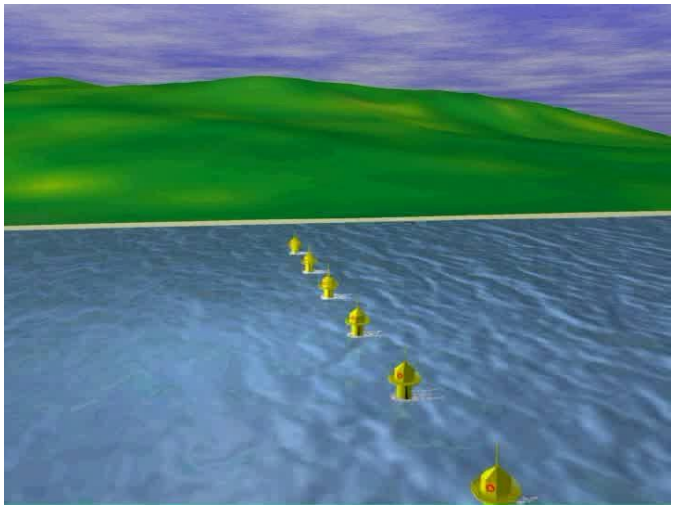
Marine turbines



Modelling marine turbines







Copyright MCT

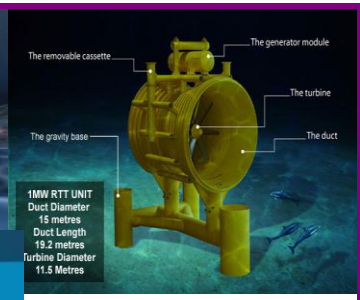
Head loss for under submerged structures (turbines):

$$\mathbf{F}_{Structures} = -\frac{\tau}{\rho h} = -\frac{S}{2hA} C_D |\mathbf{U}|\mathbf{U}$$

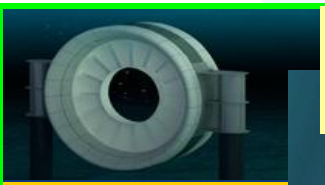
A number of possible technologies for a first real life test in Bréhat…

MCT – SeaGen (UK)

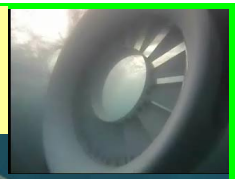
Tidal Generation (UK)

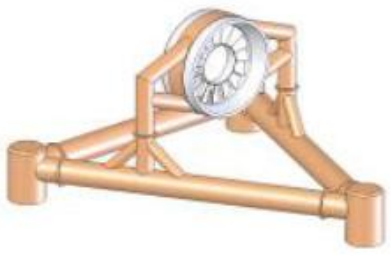


Lunar Energy - Rotech (Ecosse)



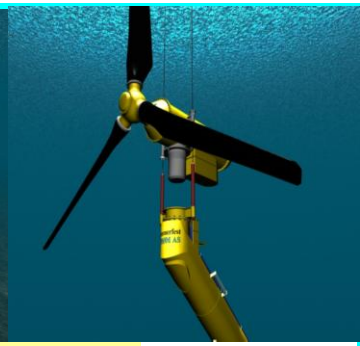
OpenHydro (IRL)





And the winner is…

OpenHydro



Hammerfest-Strøm (Norvège-Ecosse)

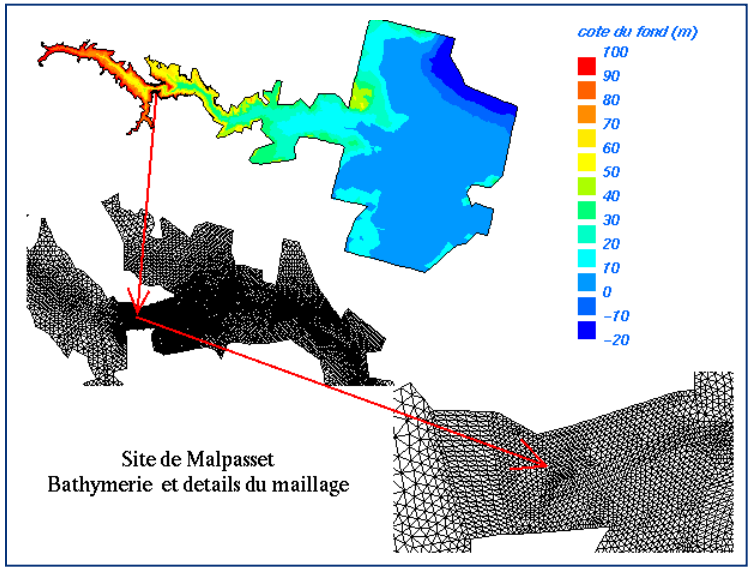
EDF R&D

Dam break and river flood modelling

Main objectives

- Conception of dams and river waterworks
- Forecasting for population safety
- Protection of industrial areas
- Damage estimation
- River basin management
- etc.

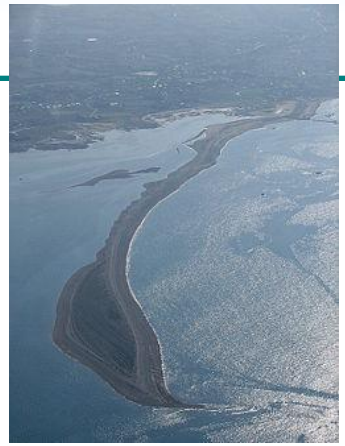
Malpasset dam break



Malpasset dam, 48 million m^3, broke on 2 December 1959, there were 433 casualties.

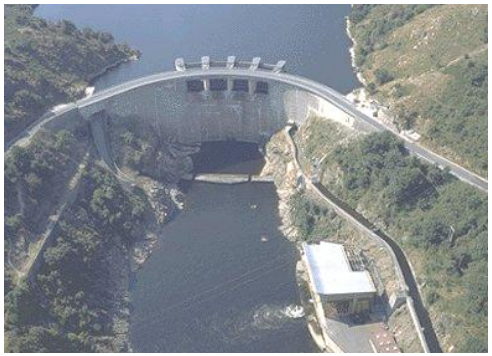
Dam for irrigation, not EDF property

The Telemac system

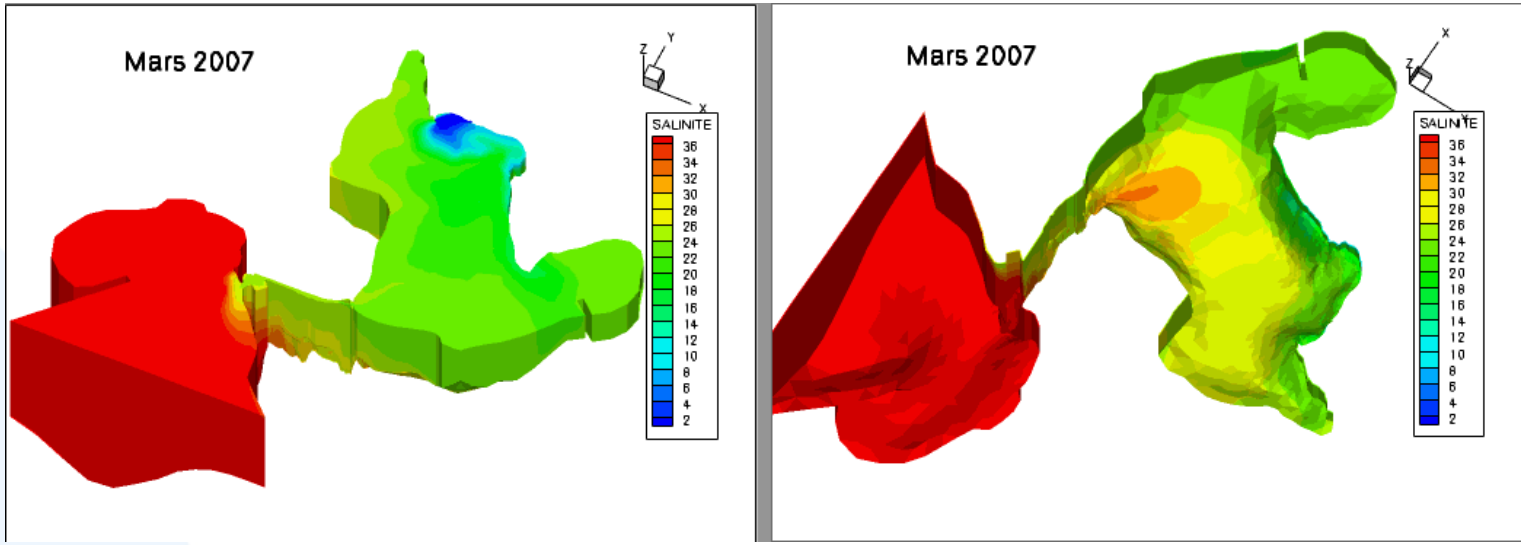


Telemac-3D

- Navier-Stokes equations
- Meshes of prisms (superimposed 2D meshes)
- Non hydrostatic 3D with free surface
- Dry zones, turbulence models
- Tracers (temperature, pollutants, sediment)



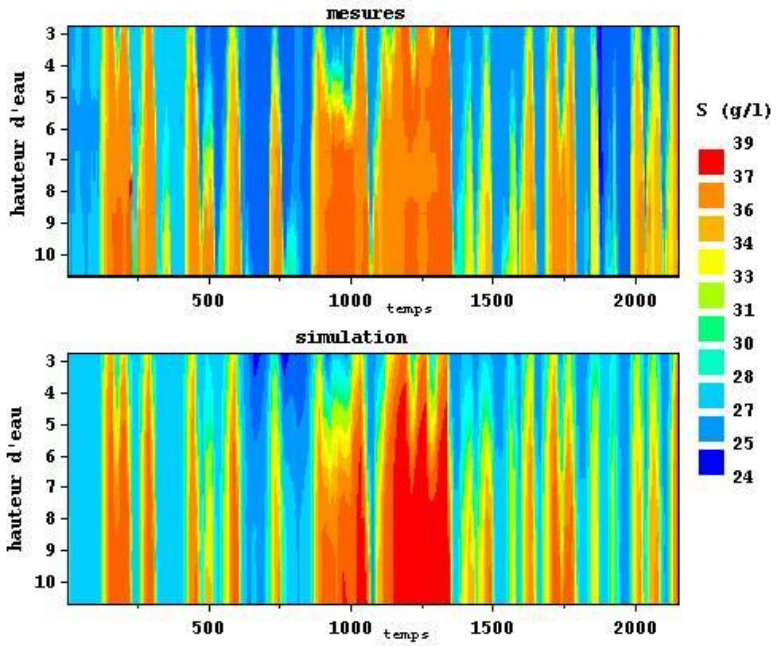
Evolution of salinity in the Berre lagoon



*** Influence of releases of Saint Chamas power plant salinity and discharges in the Caronte canal**

*** Designing scenarii for releases**

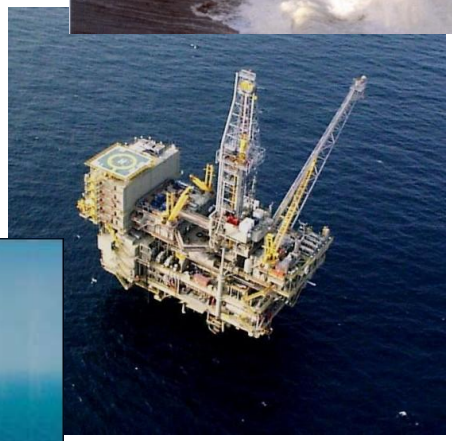
Salinity in Caronte canal between 17 November and 2 December 2005



Offshore and near-shore wave modelling

Main objectives

- Wave forecasting for coastal design
- Extreme wave prediction for navigation and people safety
- Conception of offshore platforms and wind turbines
- etc.



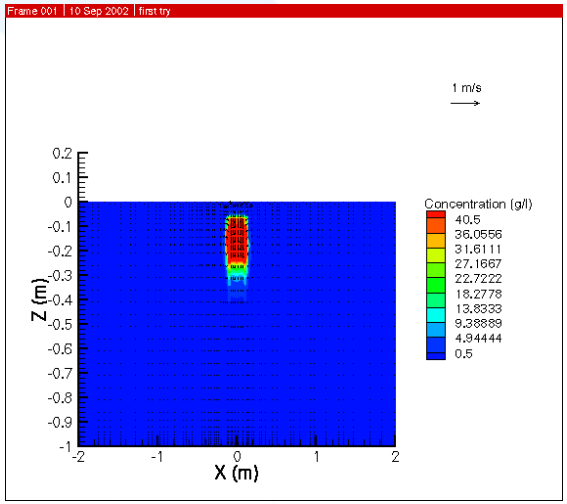
Sediment transport and morphodynamics

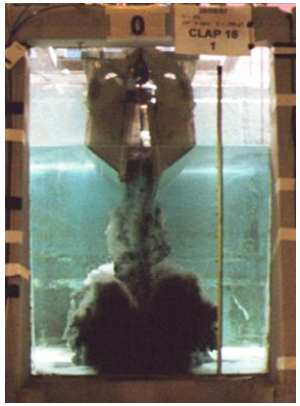
Main objectives

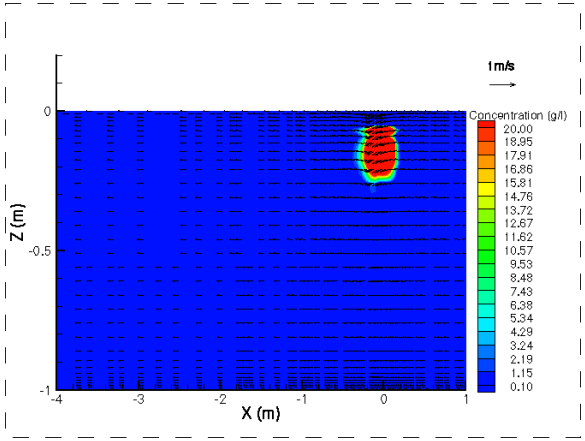
- Sand transport in oceans, coastal erosion, shoreline behaviour
- Turbidity in rivers and estuaries for water quality studies
- High mud transport rates during floods
- Morphodynamics of rivers, river meandering
- etc.

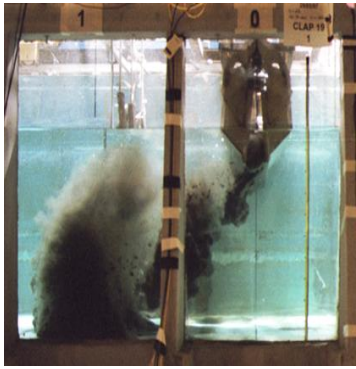
Dumping of dredged sediment

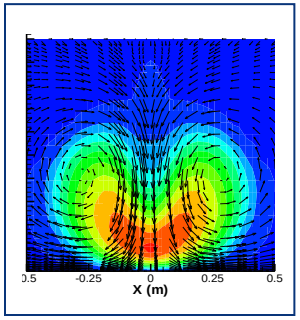
Sediments removed in harbours must be released offshore with a minimum environmental impact

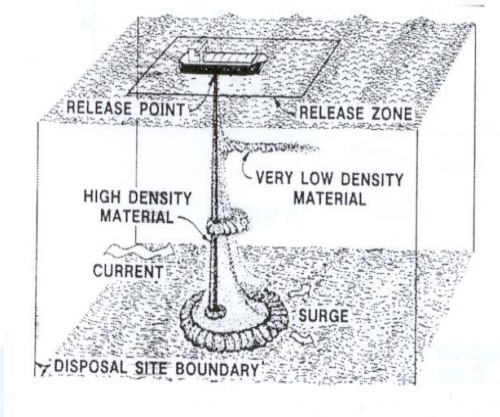


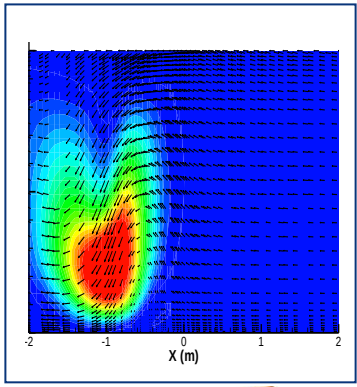




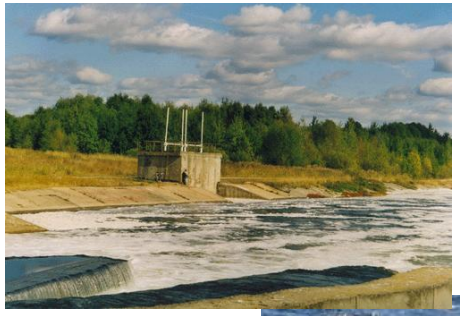








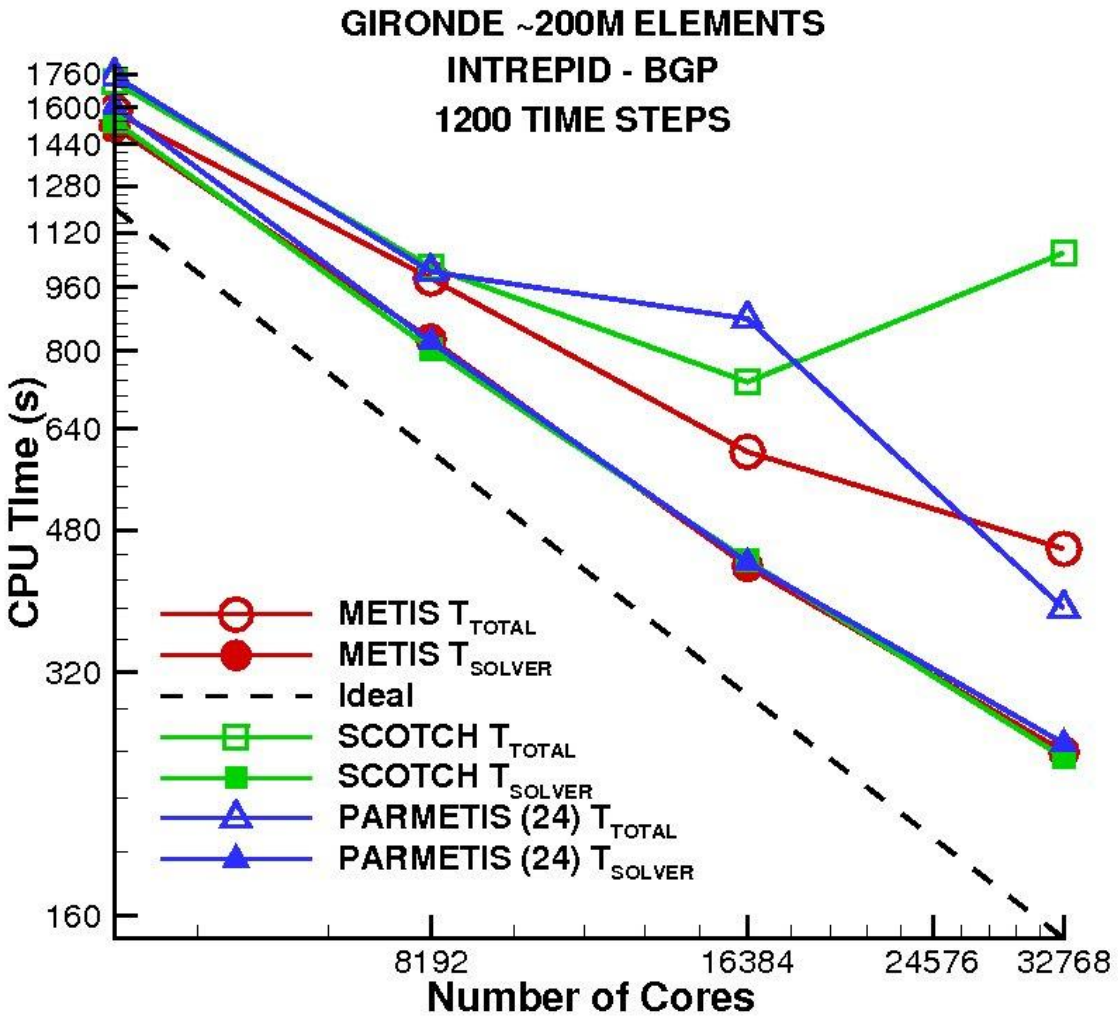
Marine and river pollution



Main objectives

- Heat transfer in oceans and rivers
- Pollutant advection and diffusion for wate quality
- Oil spill pollution
- Behaviour of algae and micro-organisms
- etc.

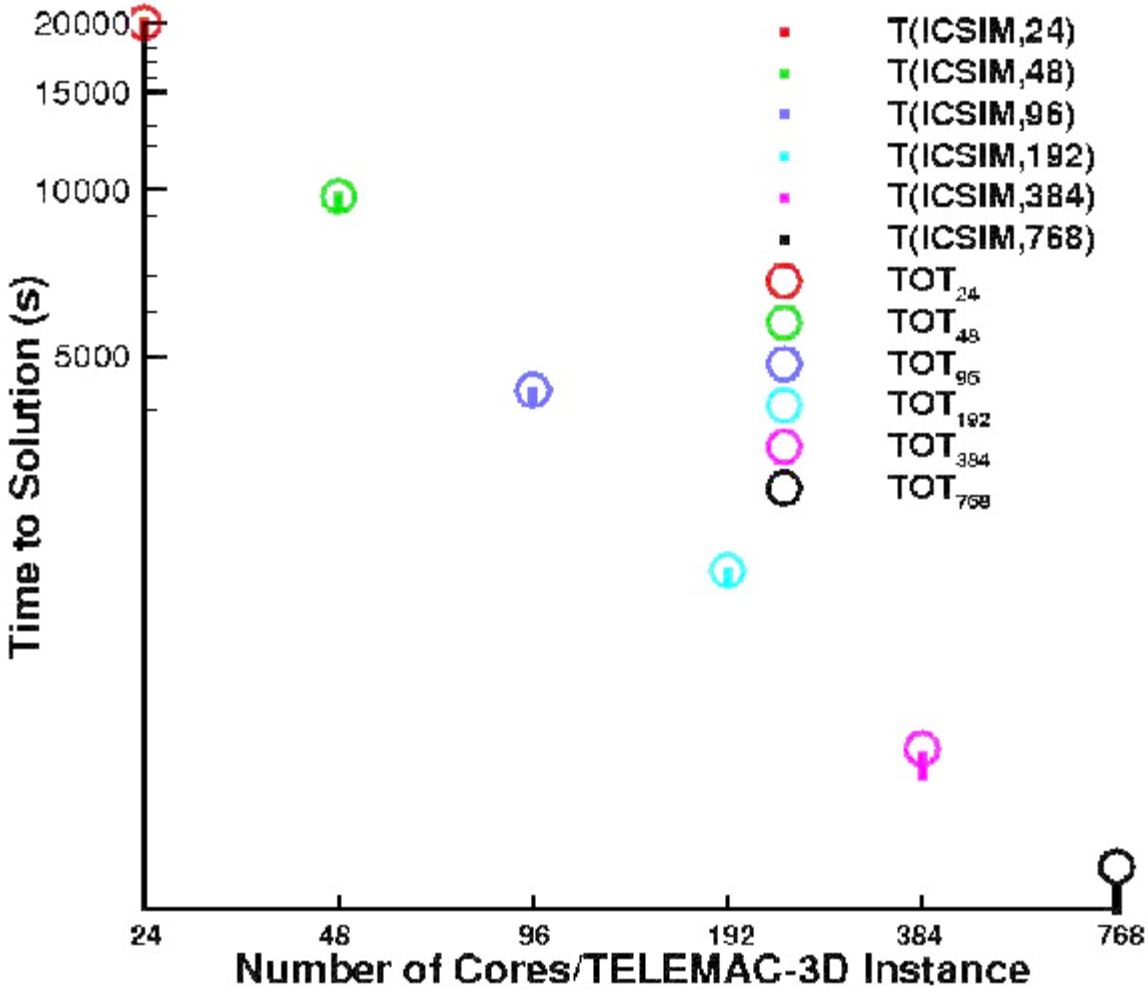
Parallel performance – TELEMAC-2D



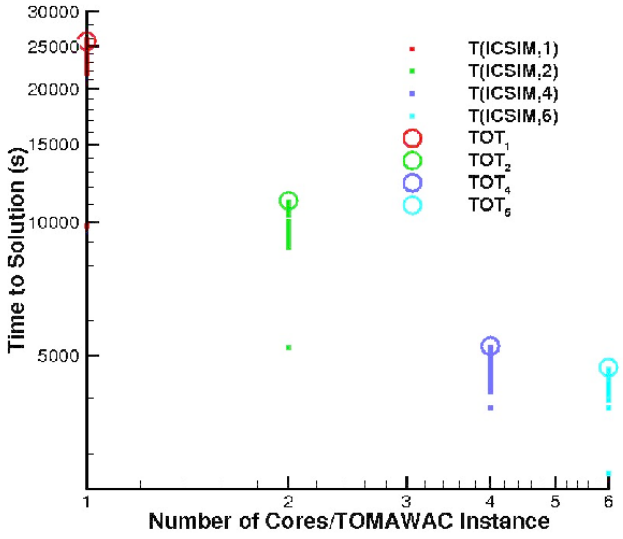
Argonne NATIONAL LABORATORY

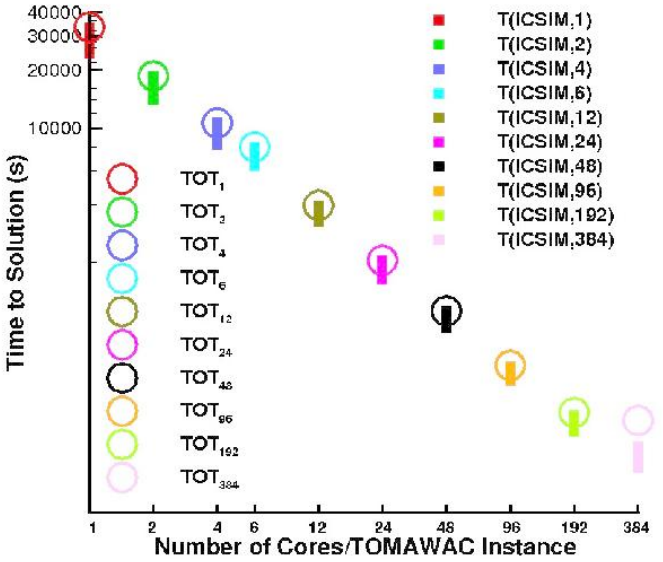
EDF R&D

Parallel performance – TELEMAC-3D



Parallel performance – TOMAWAC





Code coupling

Coupling between Hydrodynamics and Sediment Transport

Coupling between Wave Model and Hydrodynamics

Application within PRACE – 3iP: Coupling TOMAWAC - TELEMAC-3D – SISPYPHE

Published in March 2007

Wiley & sons
http://eu.wiley.com

377 pages