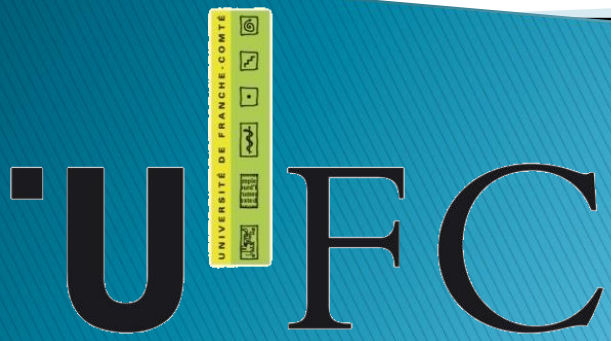
ANR AGENCE NATIONALE DE LA RECHERCHE
Work funded by ANR
ANR-07-CIS7-011-01

CPU modelling in SimGrid using dPerf

J. Bourgeois, B. Cornea
University of Franche-Comté

Cargèse, april 2010

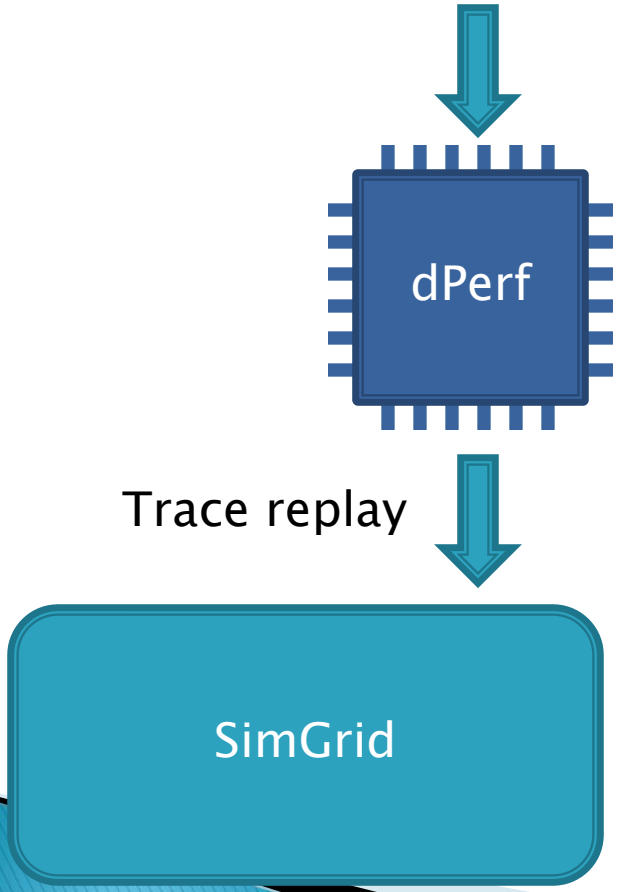


Objectives

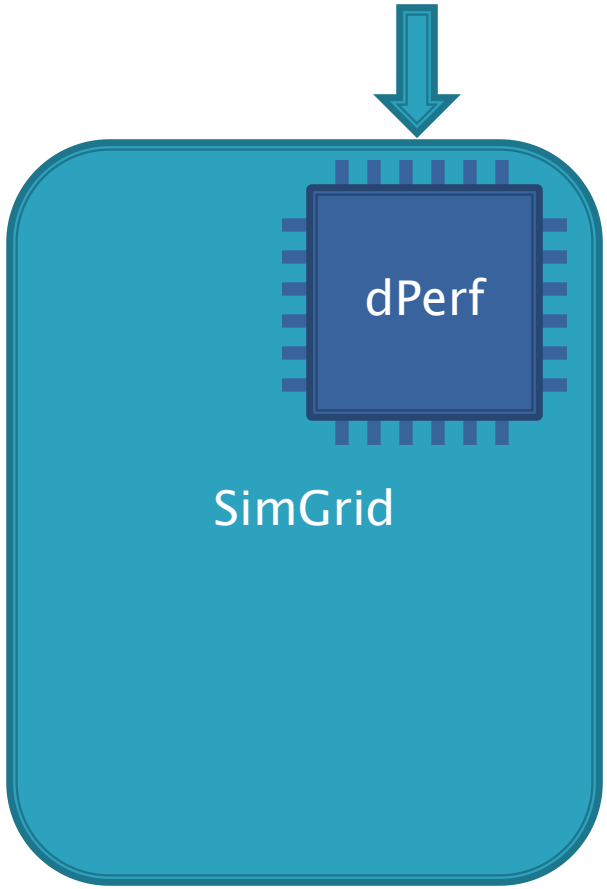
- Develop a framework to estimate performance of distributed applications
  - Written in C/C++/Fortran/Java
  - Using MPI/P2Pdc/JNGI
- Integrate computation time estimation into SimGrid
- Experiments with *real* applications in *real* conditions

Objectives

- dPerf using SimGrid



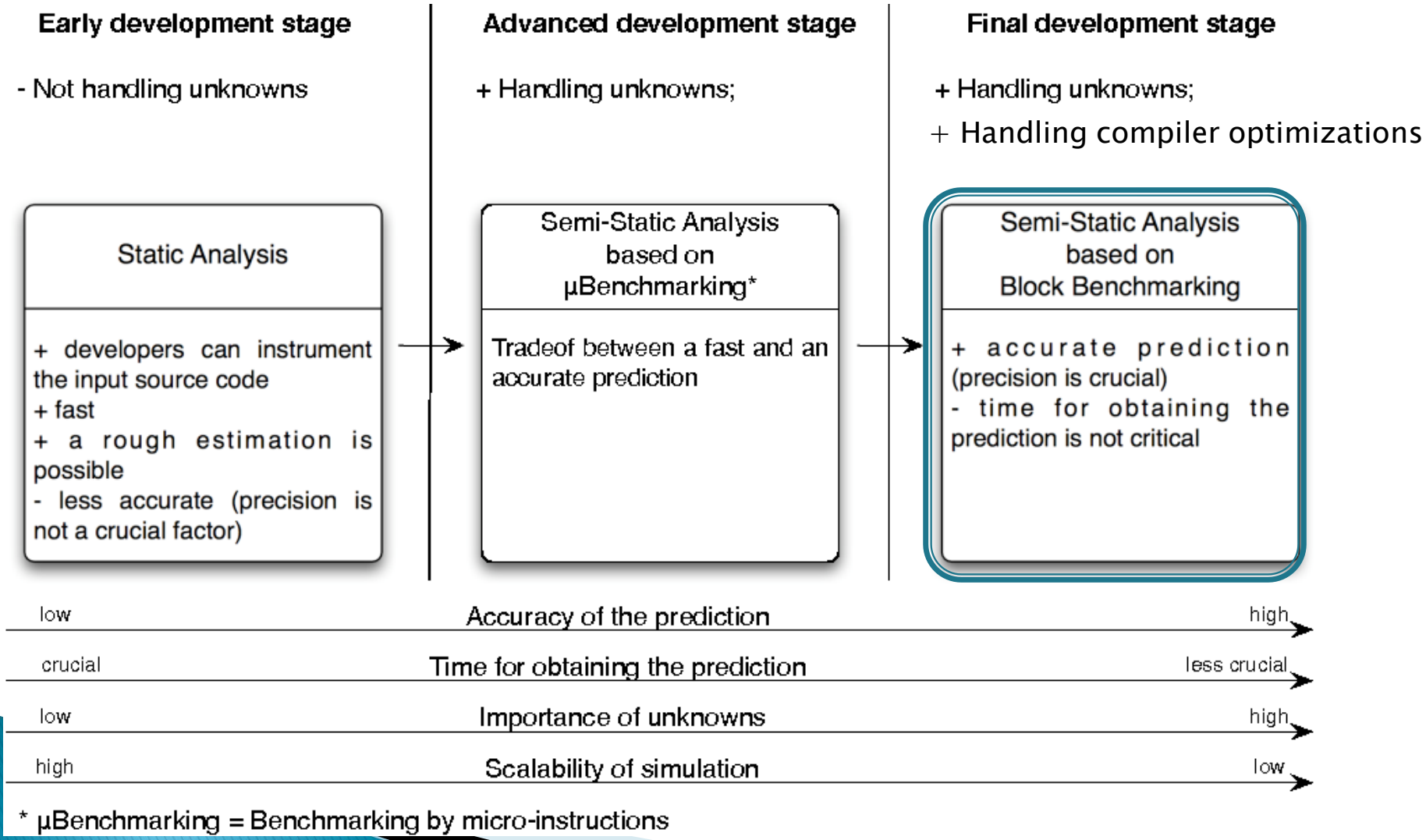
dPerf inside SimGrid



History

- 1997 EDPEPPS
  - C/PVM inside SES/Workbench
- 2001 Chronos
  - C/MPI
- 2008 P2PPerf
  - Java/JNGI inside NS2
- 2011 dPerf
  - C/C++/Fortran inside SimGrid

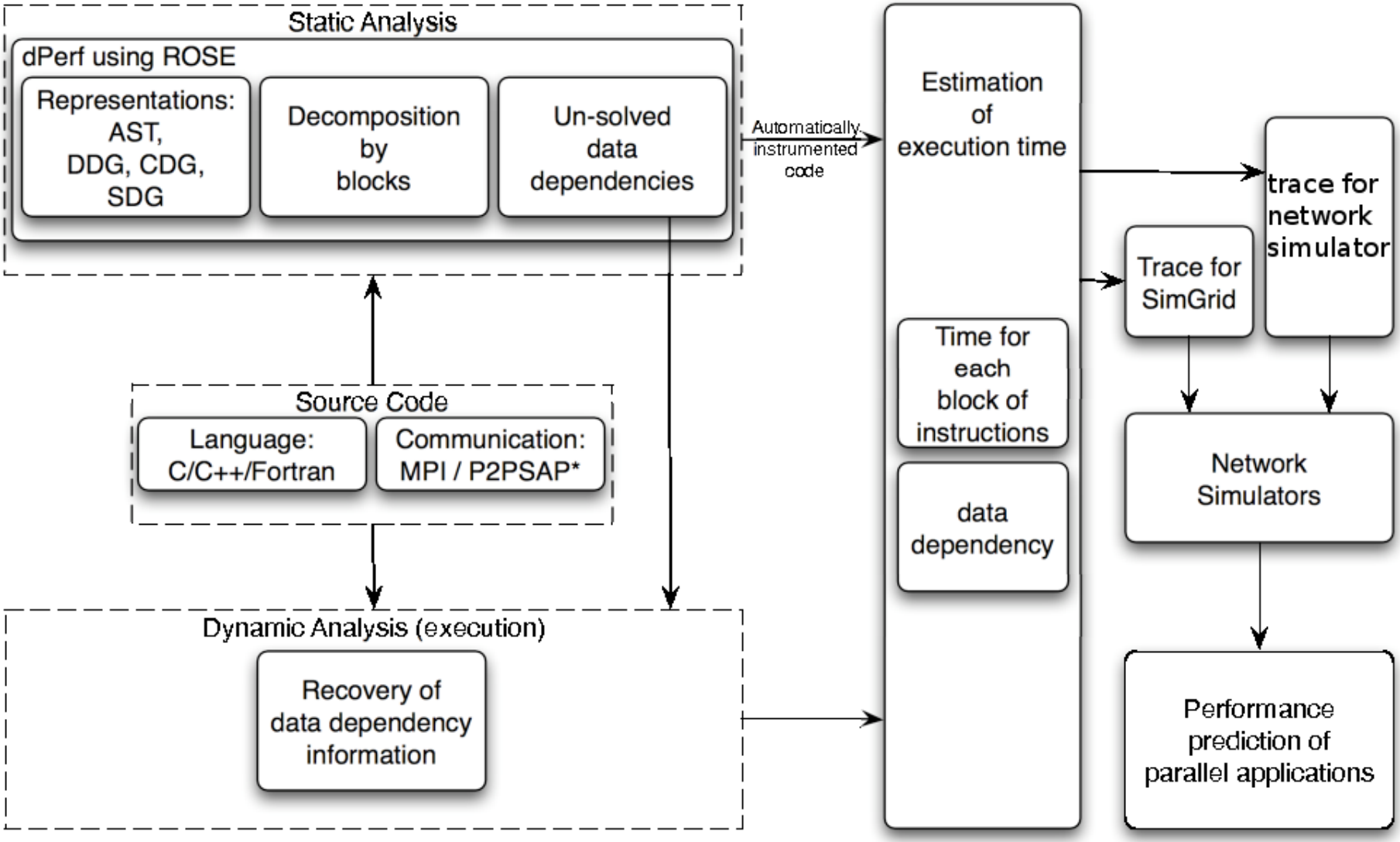
Life-cycle performance analysis



Life-cycle performance analysis

	Static	Semi-static	Block Benchmarking
Accuracy	Low	Average	Good
Slowdown	Faster	Faster	Slower or a bit faster
Unknowns	**Can't solve**	Can solve	Can solve
Compiler Opt.	Can't use	Can't use	Can use
Program parameters	Can extrapolate	**Need execution**	**Need execution**
Memory hierarchy	Can't handle	Can't handle	Handled partially
Number of computers	Can extrapolate	**Fixed**	**Fixed**
Network	Configurable	Configurable	Configurable
Program	**Simples**	Complex	Complex

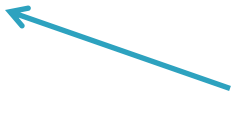
dPerf



* P2PSAP is a self-adaptive communication protocol developed by the LAAS-CNRS team, for P2P computing systems

Parametric block benchmarking

```
for(i=0;i<10000;i++)
  for(j=0;j<50000;j++)
    for(k=0;k<10000;k++)
       mult[i][j]+=m1[i][k]*m2[k][j]
```



Data dependency?

Parametric block benchmarking

```
for(i=0;i<10000;i++)
   for(j=0;j<50000;j++)
     for(k=0;k<10000;k++)
        mult[i][j]+=m1[i][k]*m2[k][j]
```

Data dependency?

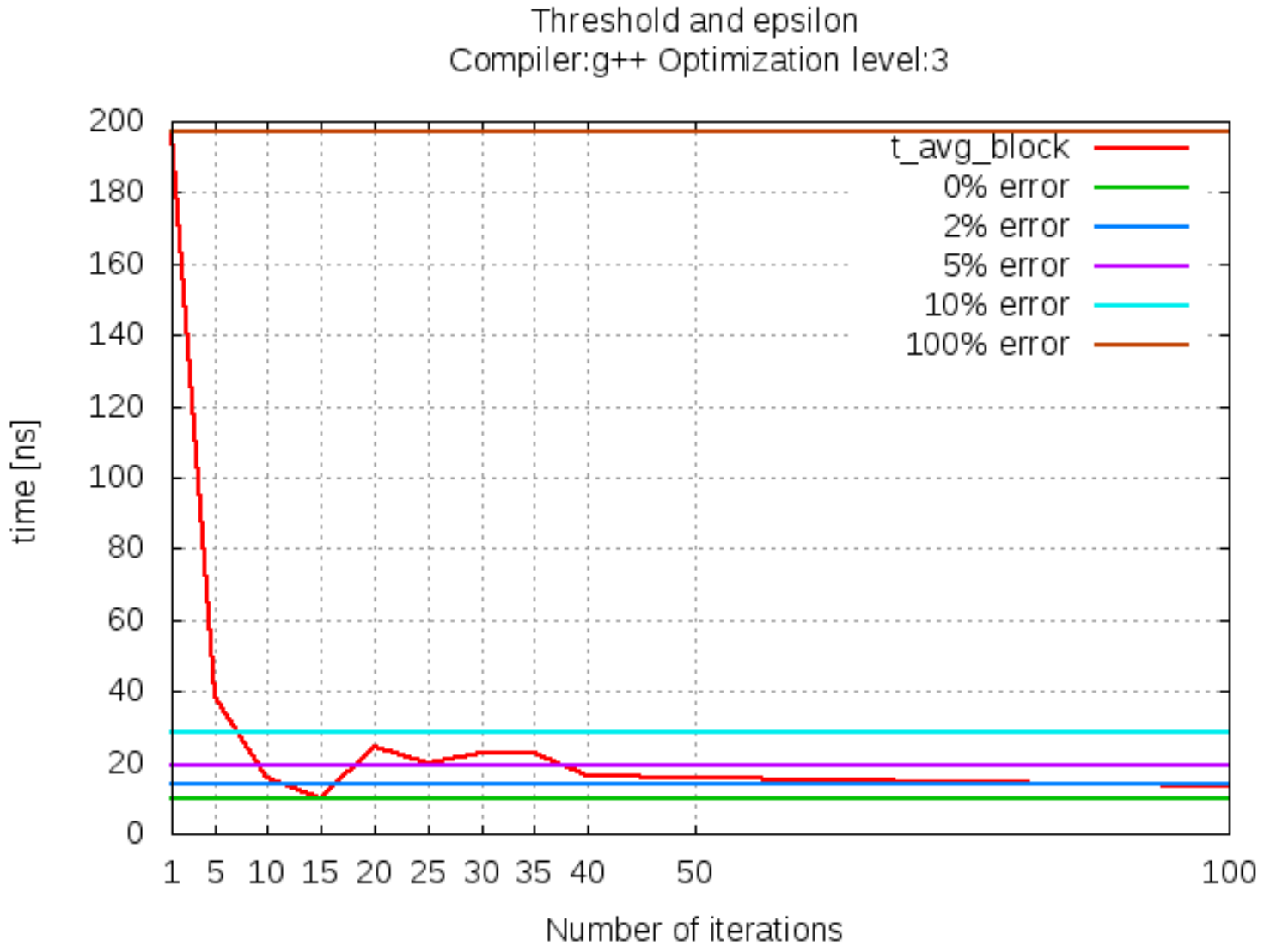
If no, the code is modified:

```
startPAPI=PAPI_get_virt_nsec();
for(i=0;i<threshold;i++)
  for(j=0;j<threshold;j++)
    for(k=0;k<threshold;k++)
       mult[i][j]+=m1[i][k]*m2[k][j];
stopPAPI=PAPI_get_virt_nsec();
```

threshold=x / t_avg_block(x)<ε with ε level of precision needed

timeBlock=[(stopPAPI-startPAPI)/(threshold^3)]*5.10^{12}

Block benchmarking



MSG-dPerf Trace File

```
[Process id] [Action] [List of parameters]

p0 compute 430564064
p0 barrier
p0 bcast 12750704
p0 send p1 389727436
p0 ssend_init p1 16632000 1409286144
p0 recv_init p1 1409286143
p0 start 1409286143
p0 allreduce 1275069469 1275069469
p0 cancel 11286143
p0 request_free 1409286143
p0 gather 1470934080
```

ns

number of bytes

MPI_request

MPI functions handled

- Already implemented by LIFC:

MPI_Gather	MPI_GatherV	MPI_Scatter	MPI_Start
MPI_Ssend_init	MPI_Recv_init	MPI_Barrier	

- Transformed at instrumentation level:

MPI_ScatterV	MPI_Startall	MPI_Waitall
MPI_Alltoall	MPI_AlltoallV	

- Undergoing implementation:

MPI_Test	MPI_Testall	MPI_Cancel
MPI_TestSome	MPI_Request_free	

Experimental testbed

- Mésocentre de l'Université de Franche-Comté
- Cluster de Calcul BullX NovaScale à base de noeuds R422E2
  - 68 noeuds bi-processeurs, Nehalem Quad Core
  - InfiniBand 20 Gb/s

Very first results

- Convection-diffusion application
  - Fortran90/MPI, ~10000 lines
  - >30 different MPI functions

- Real fortran application, **highly optimized** with MPI black magic inside! ☺

Work to be done

- Validation of experiments
- More experiments…and debugging!
- Trace file replay in SMPI instead of MSG?
- Handle multi/many cores architecture
- Simulation of P2Pdc in SimGrid
  - Wrapper?
  - MSG implementation?

- Do you need CPU timing without MPI?

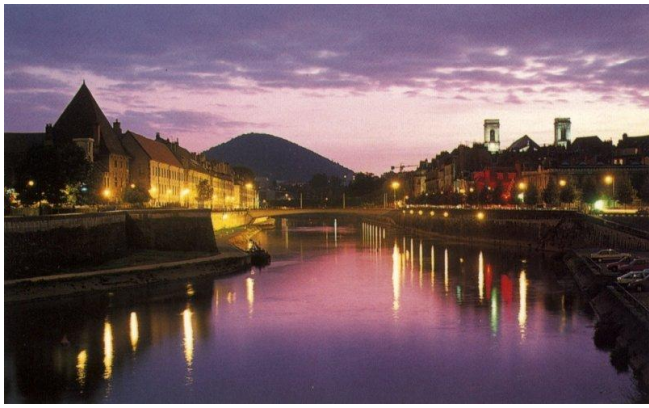
dMEMS 2010

1st Workshop on hardware and software implementation and control of distributed MEMS

Besancon, France
June 28-29th 2010

http://dmems.univ-fcomte.fr

- Network of distributed sensors and actuators
- NoC and Soc design,
- Routing and switching in embedded networks, Cross-layer design,
- Data aggregation, data fusion,
- Distributed and peer-to-peer computing, swarm intelligence,
- Distributed applications
-



	Early Registration	Standard Registration
IEEE Member Full Registration	250 Euros	300 Euros
Non-member Full Registration	300 Euros	350 Euros
IEEE Member Student Registration	150 Euros	200 Euros
Non-member Student Registration	200 Euros	250 Euros
Extra banquet ticket	60 Euros	60 Euros

Questions?