Sciomp Meeting- July 16-20 2007

High Performance Computing IBM collaborations with EDF R&D on IBM Blue Gene system

Jean-Yves Berthou – EDF R&D

Pascal Vezolle / Olivier Hess - IBM Deep Computing

Acknowledgement

EDF: Y. Fournier, C. Domain, R. Issa

IBM Research: J.P. Fassano/V. Austen, A. King, J. Sexton, C. Brooks, A. Curioni

IBM France: N. Joly, F. Bothorel

EDF R&F – IBM collaboration history

- **2005**: first collaboration with **EDF R&D** started from a BlueGene benchmark that was run in IBM France *(Europe Deep Computing Benchmark Center in Montpellier)*
  - This first series of tests gave birth to a collaboration on 2 domains: Material science with VASP and CFD with Saturne (internal code)

- **July 2006**: EDF bought a Blue Gene System (2 racks); First installation in France, hosted and administrated by IBM Montpellier PSSC

- **December 2006**: installation of 2 additional racks

- **February 2006**: The 4 racks were moved to EDF site near Paris *(Site des Renardières)*

- **July 2007:** GPFS upgrade

Four domains of interests are part of a joined EDF R&D and IBM activities (for porting, tuning and develop massively parallel applications), in order to meet "Grand Challenges".

Some Blue Gene ongoing joint efforts

- **Material Science**
  - VASP, CPMD on evaluation
- **CFD**
  - Internal codes
- **Nuclear Plant: Accident simulations**
  - CATHARE (internal) and HTC Blue Gene partition
- **Financing**
  - COINS and Mariva

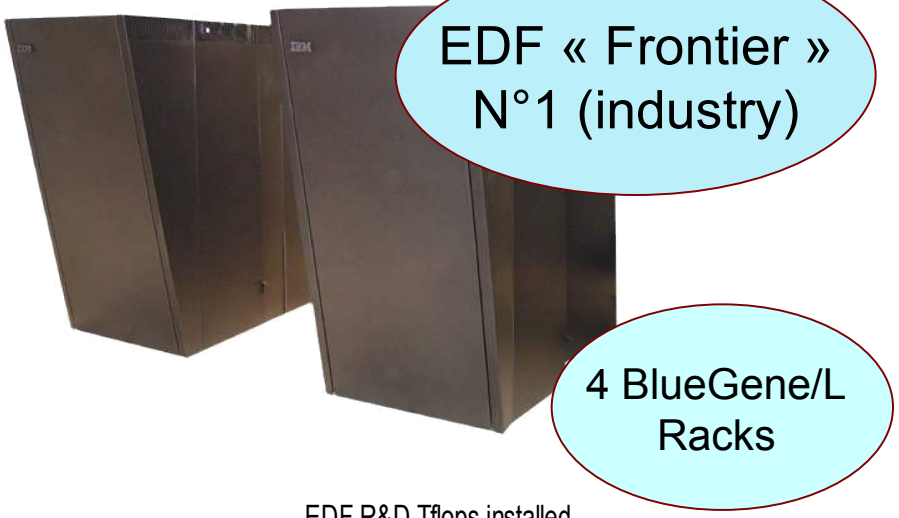
High performance simulation enables decision making

Technological breakthrough: massively parallel computing

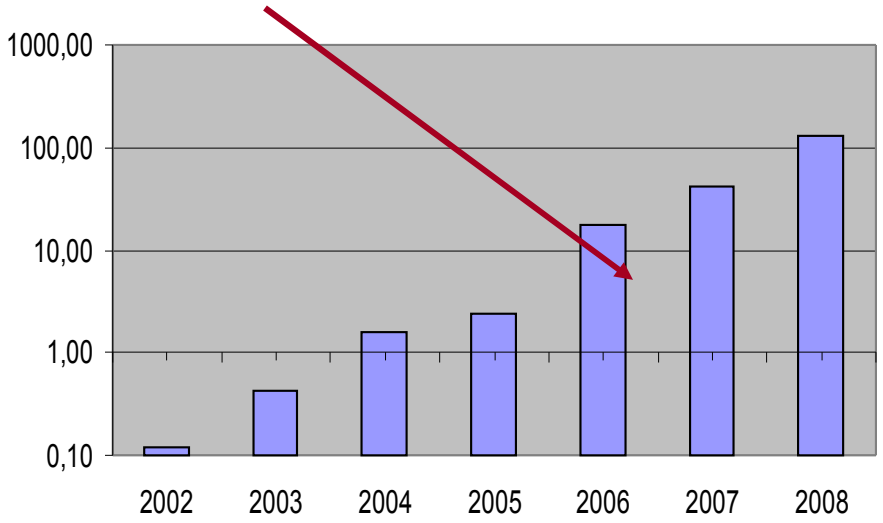
- New architectures : more than 100 000 processors
- Computing power X 1000 in the last 10 years
- modeling and applied mathematics have also made leapfrog progress.

Simulation is no longer just about understanding; it can greatly benefit the decision making processes

From understanding to decision making



EDF R&D Tflops installed

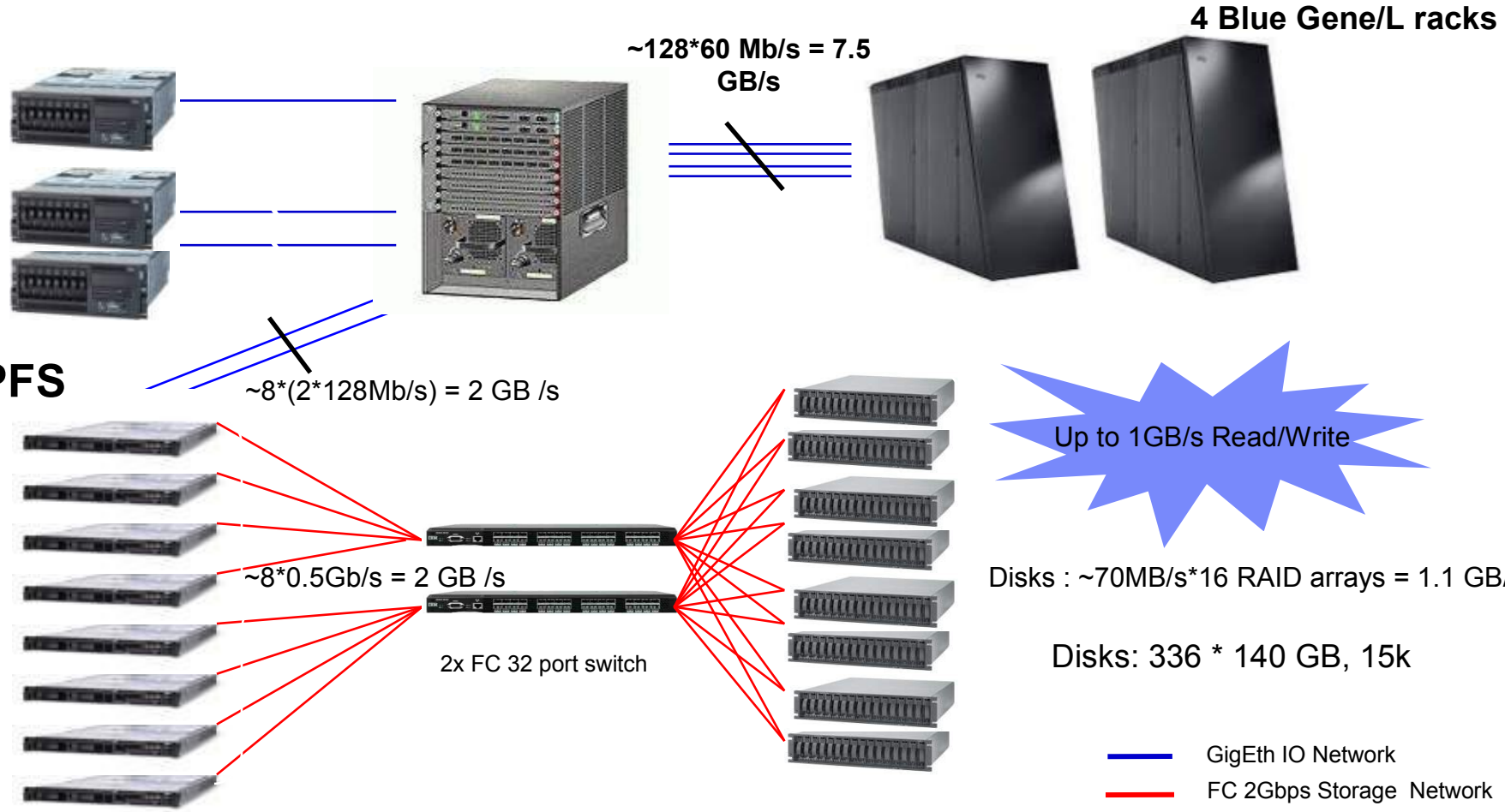


EDF Current IBM Blue Gene configuration

4 racks Blue Gene\L

GPFS file system, 8 servers 2way, 4 DS4700 => ~ 15TBytes, 1 GBytes/s

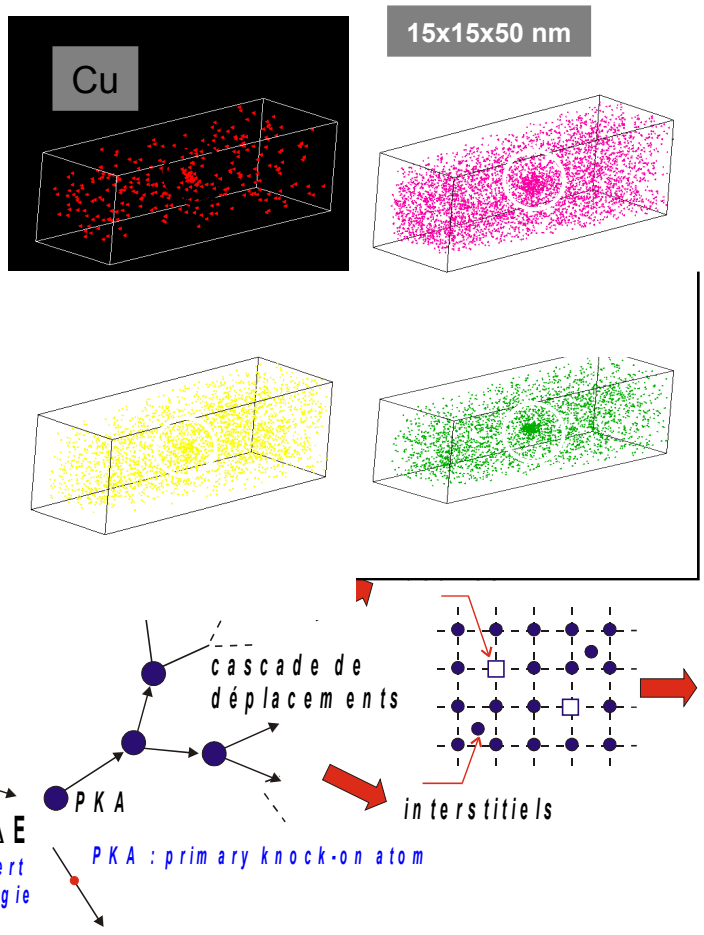
New GPFS in process: 8 servers 4 way, 6 DS4700, => ~45 TBytes, 4.5 GBytes sustained



EDF R&D and IBM collaborations

Material Science: ab initio Calculation - HPC applications for Nuclear Plant

- Irradiation damages (materials structure) :
  - Vessels
  - Internal structure
  - rod (or casing)



- HPC Applications :
  - Material Corrosion
  - Corrosion in primary circuit

IBM involvement: VASP, CPMD assessment

Molecular Dynamic

Classical molecular Dynamic

- 1960 : MD up to 100 atoms - 100 time steps
- Very simple interaction potential (LJ)

- 1990 : MD up 10.000 atoms - 10.000 time steps
- Modèle d'interaction pour métaux EAM (paires + champ local)
- 1995-97: 1.000.000 atoms - 50.000 time steps (parallelisation T3D/T3E Grenoble)
- 2007 : up to millions of atoms - qq 100.000 time steps

- 5 years perspectives :
- More than million of atoms with up tp millions of time steps.

ab initio Molecular Dynamic

- 2000 : Up to 10 atomes (T3E Grenoble)
- 2005 : Upto 250 atomes CCRT)

- End of 2006 : Up to 1000 atomes (BGL)

- 5 years perspectives (2010) :
- Several thousand of atoms
- MD with 100 à 1000 atoms - qq 100time steps

- 10 years perspectives:
- More than 10.000 atoms - 10.000 time steps

Ongoing CFD codes on Blue Gene

2 main families:

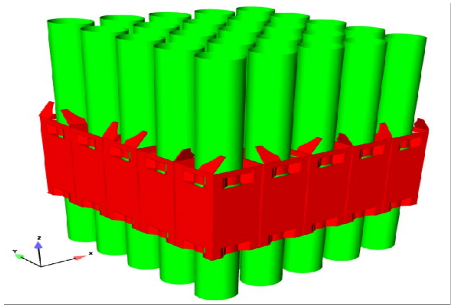
- Nuclear reactor simulation
  - Finite Volume approach, RANS
    - its co-located Finite Volume approach, it deals with any type of mesh cell and grid structure.
    - Incompressible and expandable flows with or without heat transfer and turbulence
    - Dedicated modules are available (radioactive heat transfer, combustion, magneto-hydro dynamics, compressible flows, Euler-Lagrange approach for two-phase flows, capabilities for parallel code coupling).

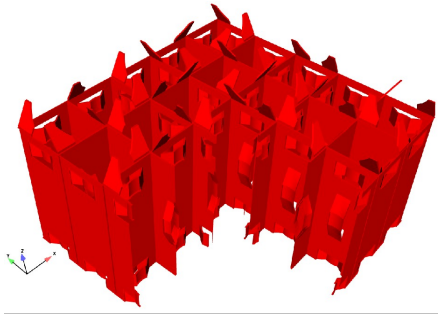
- Environment simulations (underground flows, water quality, sedimentation, dam breaking, etc …)
  - Euler, Finite Elements codes
  - Lagrangian, SPH (Smoother Particules Hydrodynamics), (dam breaking simulation),

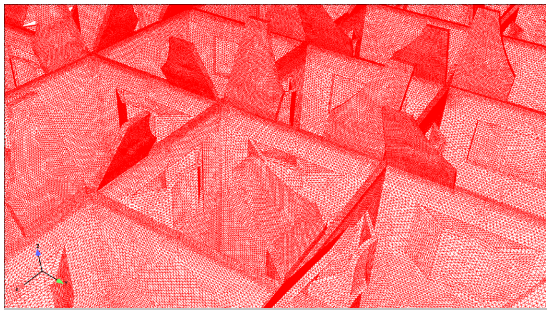
Application to fuel assemblies

100 million cells test case

- Prototype grid, with experimental data
- Meshed in 5 parts, joined by the pre-processor
- Tetrahedral mesh
  - Hexahedral mesh should be built later in the year (better for calculation convergence and quality, but much more difficult to mesh)
  - 5x5 grid rods, (twice as refined as necessary,so mesh is 2x2x2 times “too large”, but a true 17x17 grid has 11 times as many elements, so the mesh size is representative of that of a 17x17 grid rods)







•In progress, preliminary run on 4000 processors

Issues with large cases

Most difficulties associated with meshing and pre-processing

- Generation of post-processing data is done in parallel
  - Serialized by slices on rank 0, avoiding memory bottlenecks
  - Will use MPI/IO in the near future, to add performance scalability
- Pre-processing is still serial
  - Construction of ghost cells now parallel
  - 60 Gb necessary for 100 Million hexahedra, 20 Gb for 100 Million tetrahedra
  - Pre-processor can join mesh to relax meshing tool memory limits (along conforming or non-conforming boundaries), but this may lead to lower mesh quality when using non-conforming boundaries in critical areas.
- Automation of hexahedra-dominant meshing still not satisfactory
  - Conflict between automation and control: mesh quality from highly automated tools often disappointing, and tools enabling finer control of mesh quality are very time-consuming

Lagrangian – SPH applications

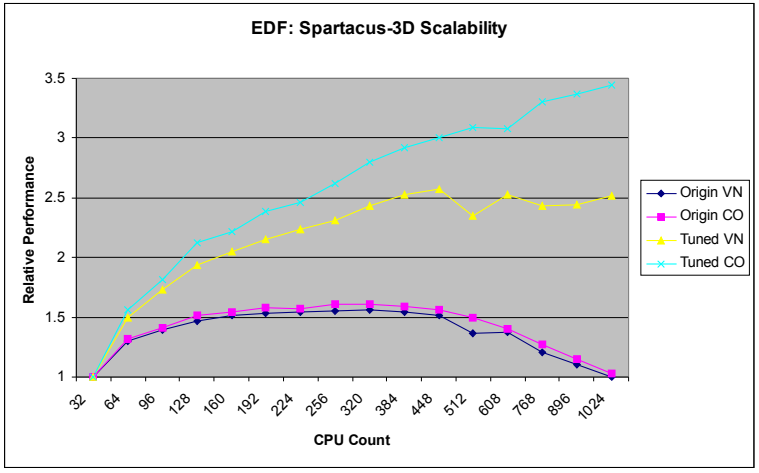
First porting to Blue Gene (realized by EDF)

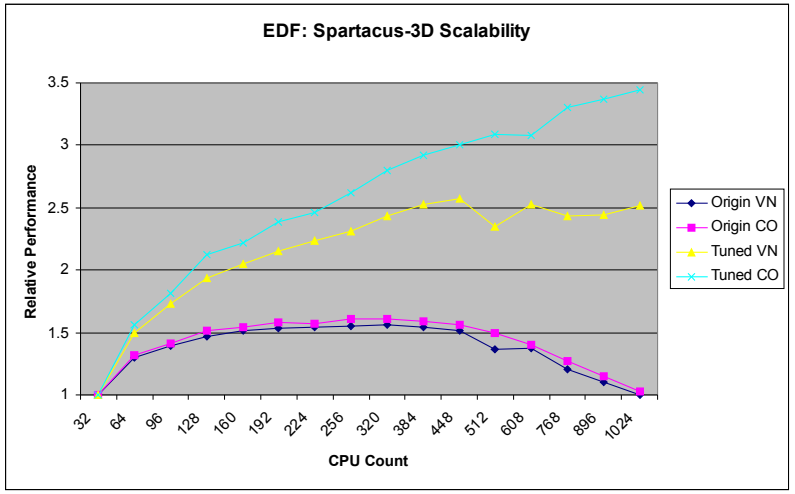
Application does not designed for a large number of procs due to memory footprint

The memory size per proc depends on the total number of particles and procs

⇒ Only 250 10^3 particles on 1024 BlueGene nodes (=> not enough particles for the scalability)

EDF & IBM on going actions: code analyze and rewriting (// data structures) to achieve 2-6 10^6 particles





IBM Blue Gene HTC capability

- **Objective: run a series of 8192 serial jobs one 4 Blue Gene racks**
  - **3-6 days of simulation time and ~ 10Tbytes of data**
    - **Serial execution time varies from 1 minute to 6 days**
    - **Total number of runs: 57344**
- **IBM is in charge to develop the interfaces between EDF's application and Blue Gene**

IBM Blue Gene HTC partition

Blue Gene/L was optimized for MPI applications

It can now handle non-MPI applications with High Throughput Computing partition

- High Performance Computing (HPC) Model
  - Parallel, tightly coupled applications
  - Programming model: MPI
- High Throughput Computing (HTC) Model
  - Large number of independent tasks
  - Programming model: non-MPI

Node Resilience

- If the block is in HTC_mode, all nodes in the reset state will be rebooted
- If the failure is not related to the network hardware, a software reboot can recover the node
  - Hardware has the ability to reset parts of the ASIC without resetting the entire ASIC
- MMCS polls every 3 seconds for nodes in the reset state
- Recover from soft failures per node such as a parity error

HTC on Blue Gene

- **HTC Launcher – resident on each BG node**
  - Listens on socket for work-requests from HTC Scheduler
  - Performs *exec(work-request)*
  - Restarts!

- **HTC Scheduler**
  - Transfers work-request to HTC Launcher collective

- ***Work-request***
  - HTC Client requests desired executable
  - BG compute node executes retrieved executable name “string” on behalf of HTC Client
  - Executable binaries stay resident until released.

Blue Gene CATHARE Implementation

EDF's constraints

- Not change the application launcher (initially implemented for PBS tool)
- Run SHELL scripts (not possible on Blue Gene)
- Each runs need a recompilation (not possible on BlueGene)

Blue Gene solution

- An interface (app_dispatcher) running on the front-end in charge to monitor job submissions through PBS wrappers and to reschedule them to BlueGene HTC partition or Linux systems depending on the type of the job
  - SHELL scripts and compilations are dispatched to a Linux Power system controlled by Loadleveler Tivoli Batch Scheduler.
  - BlueGene jobs are sent to htc_dispatcher using IPC structures for performance (shared segment and Message Queue)
- The Blue Gene jobs are managed by a standard HTC implementation

Advantages

- interoperability between Blue Gene and other systems; collaborative workload able to use all the Hardware resources
- No independency on a specific Batch tool
- Capability to include on a grid environment
- Standard BG HTC implementation

How does it works (job submission only)