

Sciomp14- May 19-23 2008

# **EDF R&D and IBM collaborations on Complex CFD applications using Blue Gene L and P**

Pascal Vezolle - IBM Deep Computing, [vezolle@fr.ibm.com](mailto:vezolle@fr.ibm.com)

Jean-Yves Berthou – EDF R&D

## Acknowledgement

**EDF:** Y. Fournier (Saturne code), R. Issa (Spartacus), M. Boucker (Neptune code), Emile RAZAFINDRAKOTO (Telemac 3d), Estelle Desroches (Telemac 3d), Ange Caruso

**IBM Rochester:** Thomas Budnik, Mark Megerian (HTC involvement)

**IBM Europe:** David Latino (Spartacus)

# EDF groupe Overview

## Permanent objectives

- guarantee safety
- improve performances/costs
- maintain assets

## Changing operating conditions

- face unexpected events, ageing issues, maintenance
- improve performance through new technologies, new operating modes and system-wide optimization
- adapt to evolving set of rules (safety, environment, regulatory)

## In-house technical backing

- expertise: strong Engineering and R&D Divisions
- physical testing and simulation are key tools from the outset

## First Electricity maker and provider in Europe

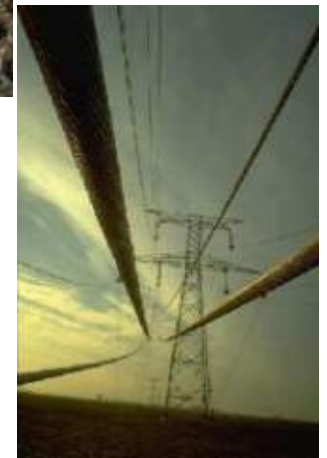
Employees in the world : 161.560

### R&D Staff

2,000 staff, 30% women

300 PhDs and 200 doctoral students

150 researchers teach at universities and engineering schools



## 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 + 2 racks); first installation in France, hosted and administrated by IBM Montpellier PSSC up to 02/2007,
- **December 2007**: EDF ordered 8 BGP racks, installed in 2008 and hosted by IBM
- **Several domains of interests are part of IBM technical support:**
  - Material science
  - Financing & Project management
  - **CFD**

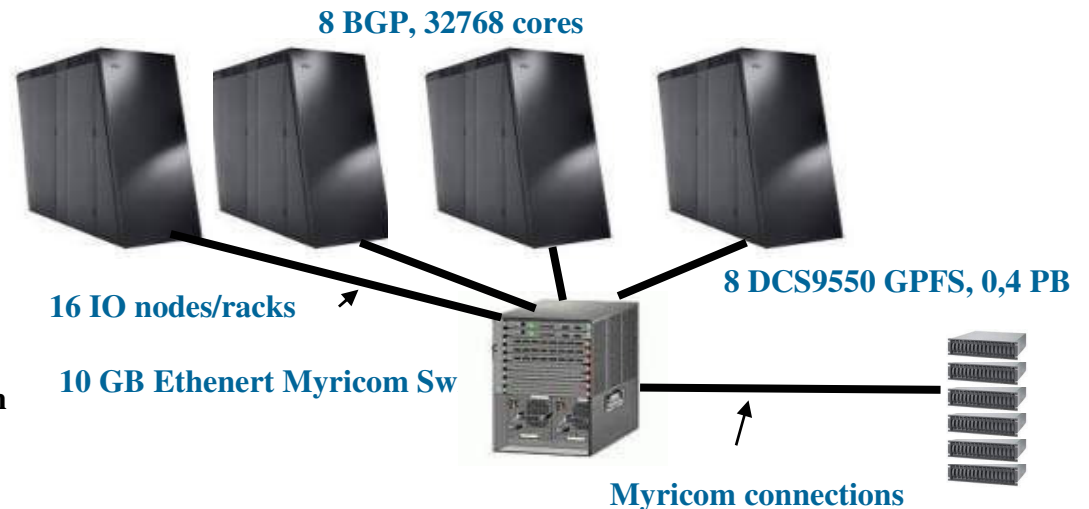
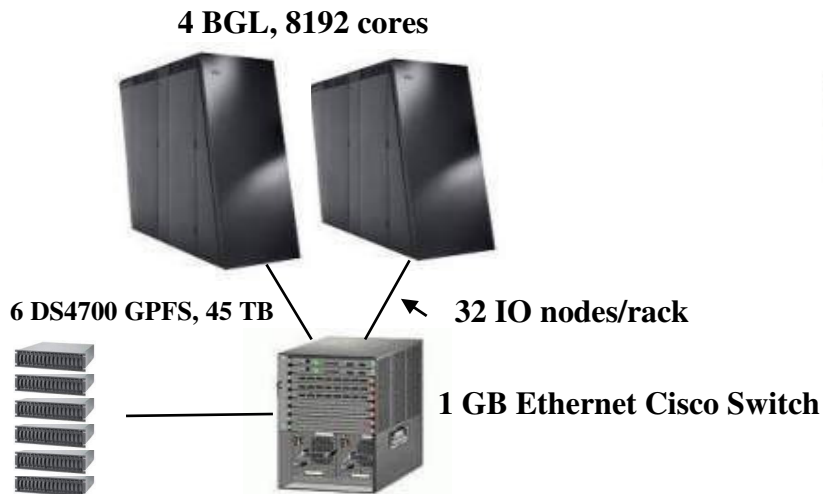
## 2 Independent Blue Gene configurations

### Blue Gene\L Configuration

- 4 racks, 4096 nodes, 8192 cores
- 32 IO nodes per rack
- 1 GBytes Ethernet Cisco Switch
- 2 System p Frontend nodes
- GPFS (~45 TBytes, ~4.5 GB/s sustained)
  - 8 System p NSD servers
  - 6 DS4700 (Scsi Disks)

### Blue Gene\P Configuration

- 8 racks, 8192 SMP nodes, 32768 cores
- 16 IO nodes per rack
- 10 GBytes Ethernet Myricom Switch
- 2 System p Frontend nodes
- GPFS (400 TBytes, ~ 8 GB/s sustained)
  - 16 System x NSD servers
  - 4 DCS9550 (DDN technology, SATA Disks)
- *IO configuration (capacity and bandwidth) will be doubled in 2009*



## IBM role

### 1. Porting/Tuning and Scaling applications to Blue Gene

- Improve scalability (communication scheme, load balancing, memory size control), serial performance (double Floating point, Compiler options), IO performance (parallel IO, GPFS tunings).
- mixed mode approach Thread/MPI on BGP

### 2. Proposing and providing a complete production environment

- How to deal with multiple-fields/applications simulations on Blue Gene (MPI and serial), involving several codes and partners
- How to run a mixed workload including BG and an external system
- How to manage TBytes of data (IO & visualization capabilities)

While keeping in mind that the applications can run also well on a standard Linux Cluster  
The whole lot being conducted and approved by EDF

**\* IBM technical Resources: locally in France (2-3 people) + US Lab Watson&Rochester support (2-3 people)**

# Ongoing CFD codes on Blue Gene

2 general, complex and complete integrated CFD families including **several difference codes** and a **workflow environment**

## ➤ 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).
- 2 main codes: Code\_Saturne (single-phase flows) and NEPTUNE\_CFD (two-phase flows)
- Code\_Saturne CFD package is Open Source: [//rd.edf.com/code\\_saturne](http://rd.edf.com/code_saturne)

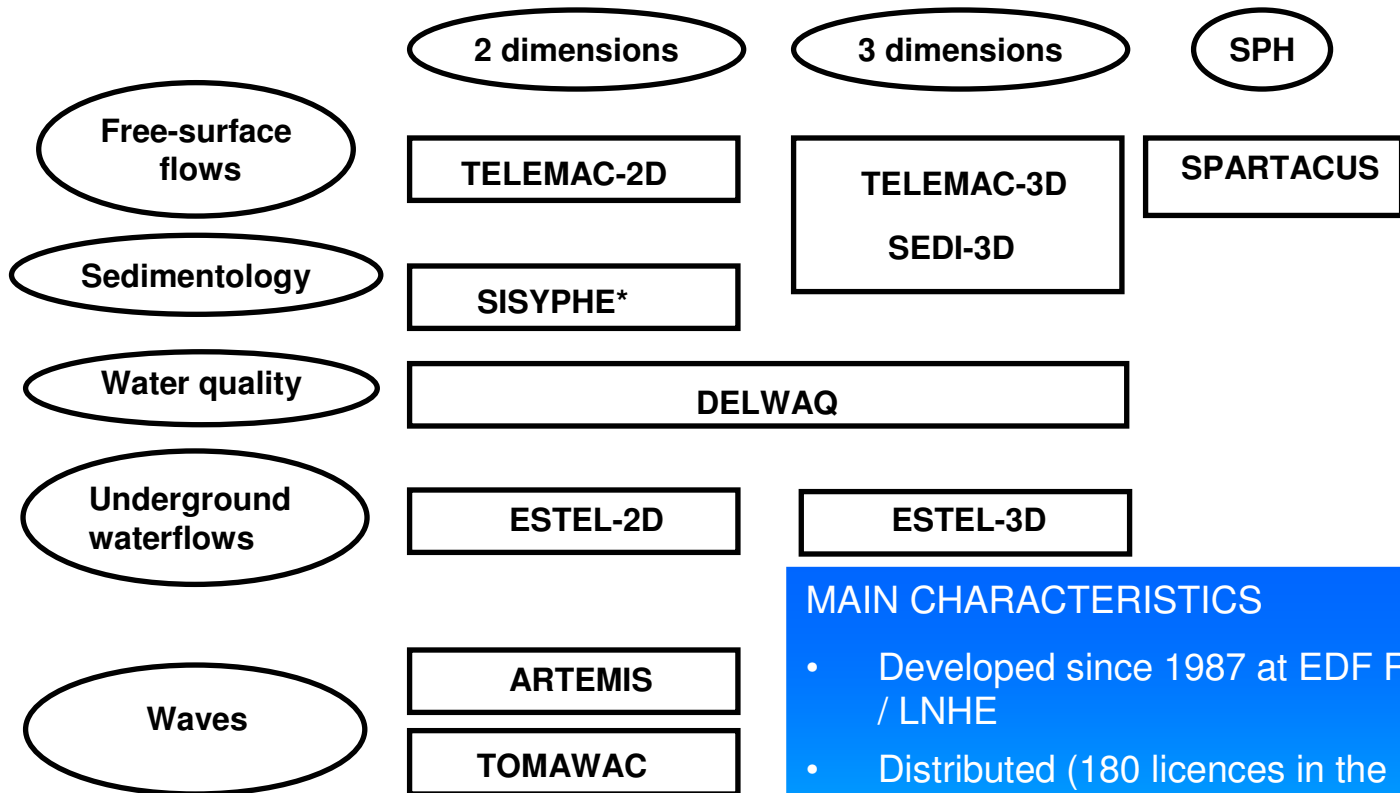
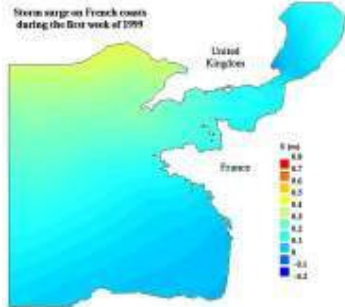
## ➤ Environment simulations: underground flows, water quality, sedimentation, dam breaking, etc ...

- **Finite Elements approach**, Euler (Telemac code), available through license, more than 100 user over the world
- **Lagrangian, SPH** (Smoother Particules Hydrodynamics), (dam breaking simulation), strong deformation (Spartacus code)

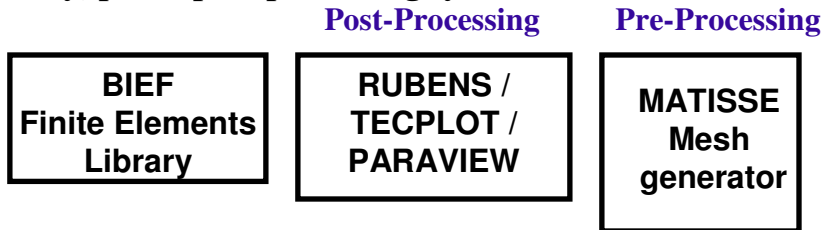


Example of Complexity 1: more than 9 different codes

Finite Elements family: The hydro-informatics system **TELEMACH**



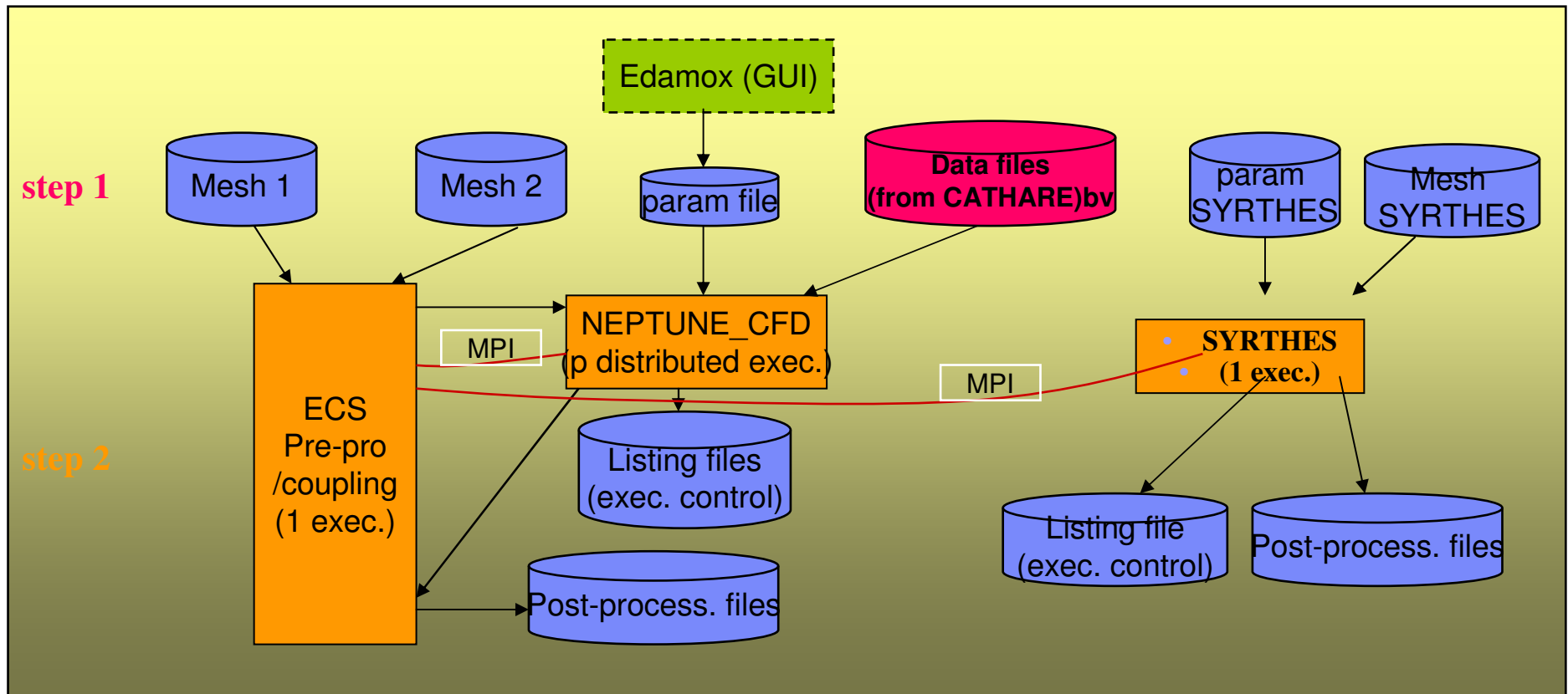
Common library, pre et post-processing systems



- MAIN CHARACTERISTICS**
- Developed since 1987 at EDF R&D / LNHE
  - Distributed (180 licences in the world)
  - Dedicated to the modelling of the main environmental phenomena in hydraulics (tides, floods, waves, marine pollutants, underground flows, etc.)

## Example of Complexity 2: chaining and coupling simulations

Focus on NEPTUNE\_CFD (Finite Volumes/two-phases, MPI) /  
SYRTHES coupling step (MPMD mode)



## Massively approach challenges for CFD (real complex simulations !!) on Blue Gene

Complex industrial CFD simulations on Blue Gene don't only mean Porting/Tuning and Scaling BUT also Compatibility, Integration, Validation and Collaboration

### 1. Standard Massively parallel standalone implementation challenges on Blue Gene

- ▶ Most difficulties associated with meshing (quality control) and pre-processing (partitioning and load balancing)
- ▶ Main limitation on Blue Gene for large simulation is the memory per core
- ▶ Time step control is mandatory to control the overall elapsed time (multigrids)
- ▶ Serial optimization and mixed mode implementation can provide a extra level of scalability
- ▶ Visualization and IO management can become critical

### 2. Multiple-physics and applications simulation (coupling and chaining)

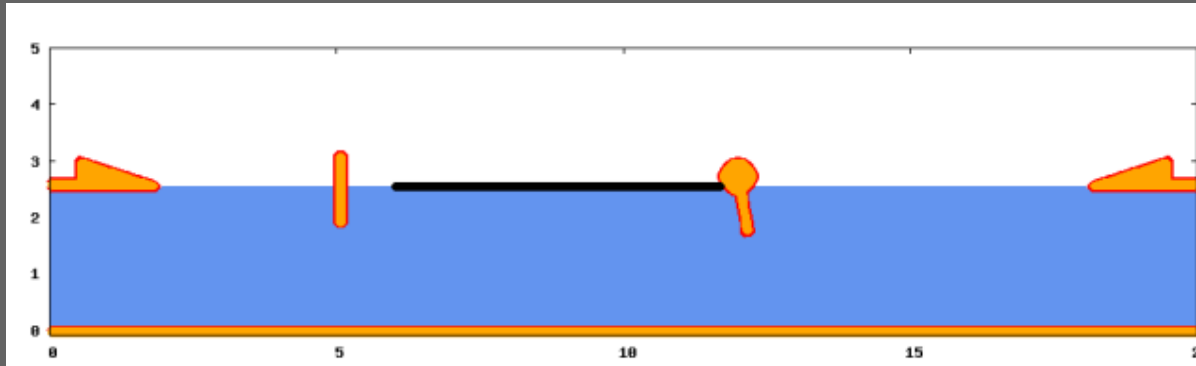
- ▶ Solution on Blue Gene\P: HPC and HTC partitions + socket communication
- ▶ Blue Gene integration in the industrial and workflow environment

### 2 examples of IBM Involvements

1. **SPH // codes (Spartacus):** code rewriting for massively parallel + validation of the visualization environment based on Blue Gene execution.
2. **Thermal-hydraulics mutlicodes simulation project on Blue Gene with serial and MPI execution:**s code porting + design of the runtime environment including HTC and HPC

# SPARTACUS-3D

3D-lagrangian modelling of complex free-surface flows



## Marine and coastal field

- Design of protection works, water intake and release works for thermal powerplants
- Design of structures (piles, foundations) for windfarm, marine current turbines

Fast-dynamics water flows and complex free surface

## River field

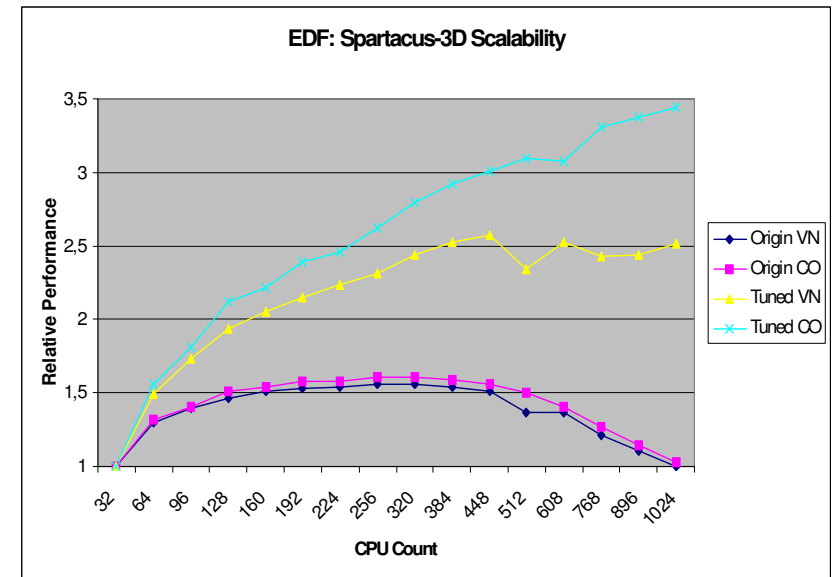
- Dam spillway crest conception
- Optimization of fish migration devices
- **Multiphase**
- **Complex geometry**

# SPARTACUS-3D performances concerning parallelism (in 2007)

## First BGL assessment tests:

- Calculation limited to 250 000 particles **due to memory per process and communication scheme**

- Decrease of the speed-up when we use more than 32 processors



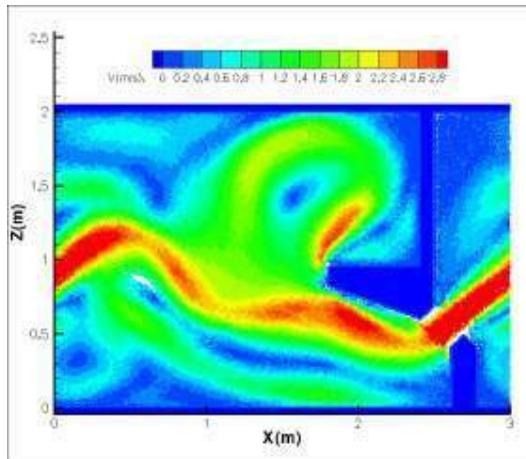
**Simple change on communication scheme**  
**removal of needless messages**

Collaboration with IBM team – Optimization performed by an IBM expert in massively parallelization methods (IBM) and an expert in hydraulics (EDF)

# What we hoped thanks to HPC

## *Fishes migration device*

- 6 millions de particles
- 10 seconds of physical time
- 50 days on 32 processors cluster



→objective : about 8 hours CPU time with BG/L

## *Dam spillway*



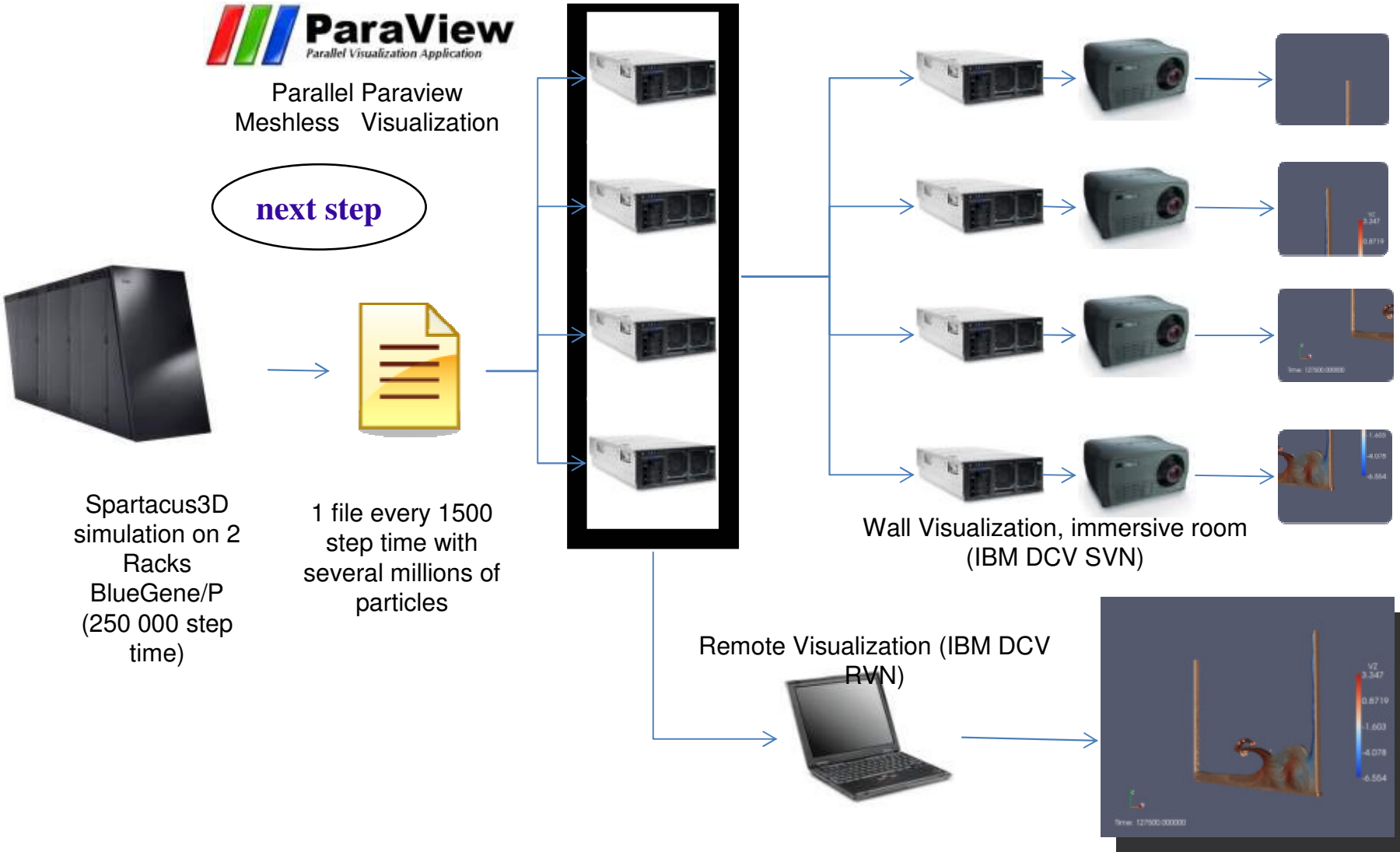
- 10 millions de particles
- from 1 to 5 minutes of physical time
- 300 days on 32 processors cluster

→objective : from 2 to 10 days CPU time on BG/L

## **IBM works to provide the best code for Blue Gene**

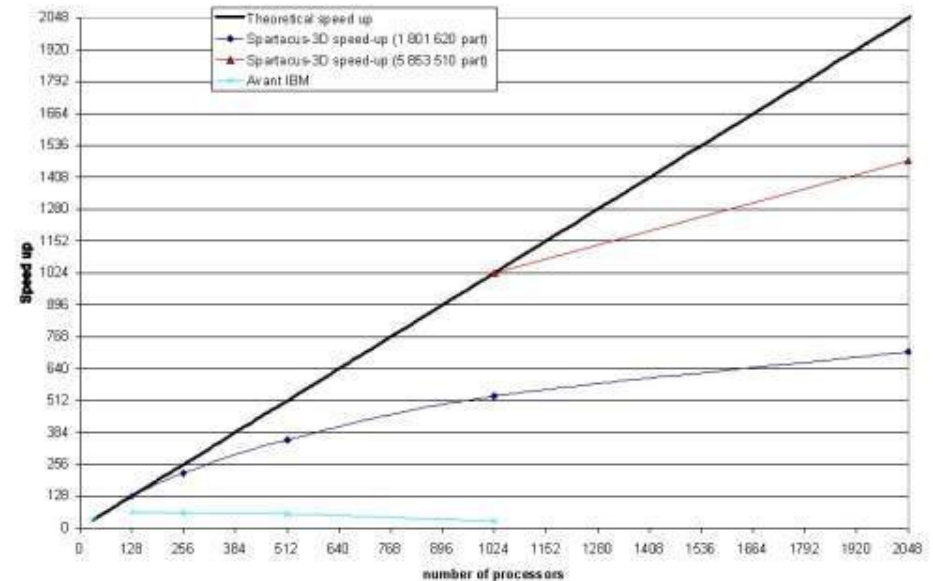
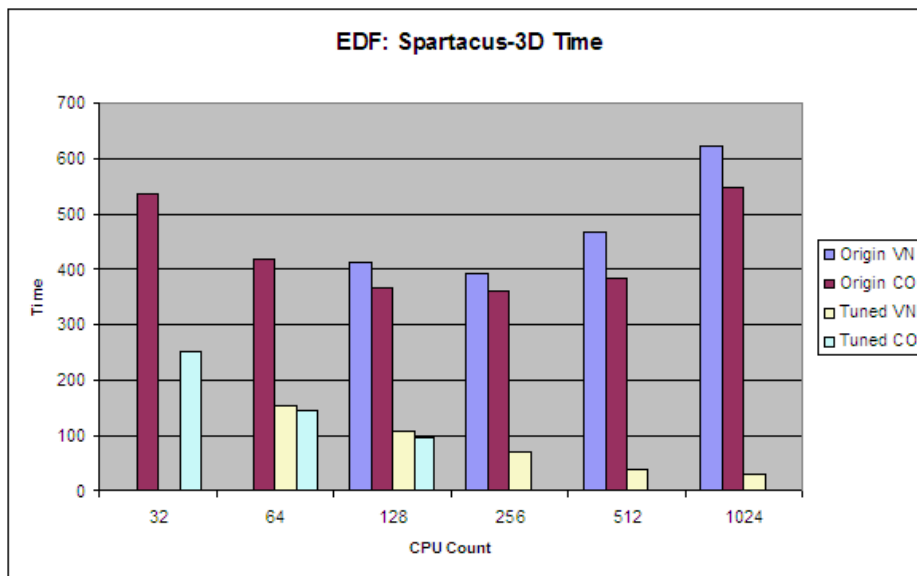
- **Fully rewriting of the communication scheme and redesign of the memory management**
  1. Reduction of the memory footprint per process (removal of the global arrays)
  2. Optimization of the communication, replacement of global collectives by neighbor point-to-point non blocking communications using RDMA features of the BG Torus network.
  3. Parallelization of the input/output files using HDF5 library
  4. Introduction of a mixed OpenMP/MPI implementation
  
- **+ Collaboration and validation on the visualization step for large models**
  1. installation in IBM site of the DCV Prove of Concept system, validation of RVN (Remote Visualization Network) through local tests
  2. New output files format implementation in Paraview Meshless realized by a EDF's partner (CSCS Manno, Switzerland)
  3. Realization of 3 movies with 250K, 2M and 5M of particles basd on BGP executions

# IBM Visualization solution



## Blue Gene Performance & Results

- From 250 000 particles on the total BG/L to several millions of particles BG/L racks (>20 millions by BG/P rack)
- Important increase of the speed up. The software is 20 times faster than before on 1024 processors

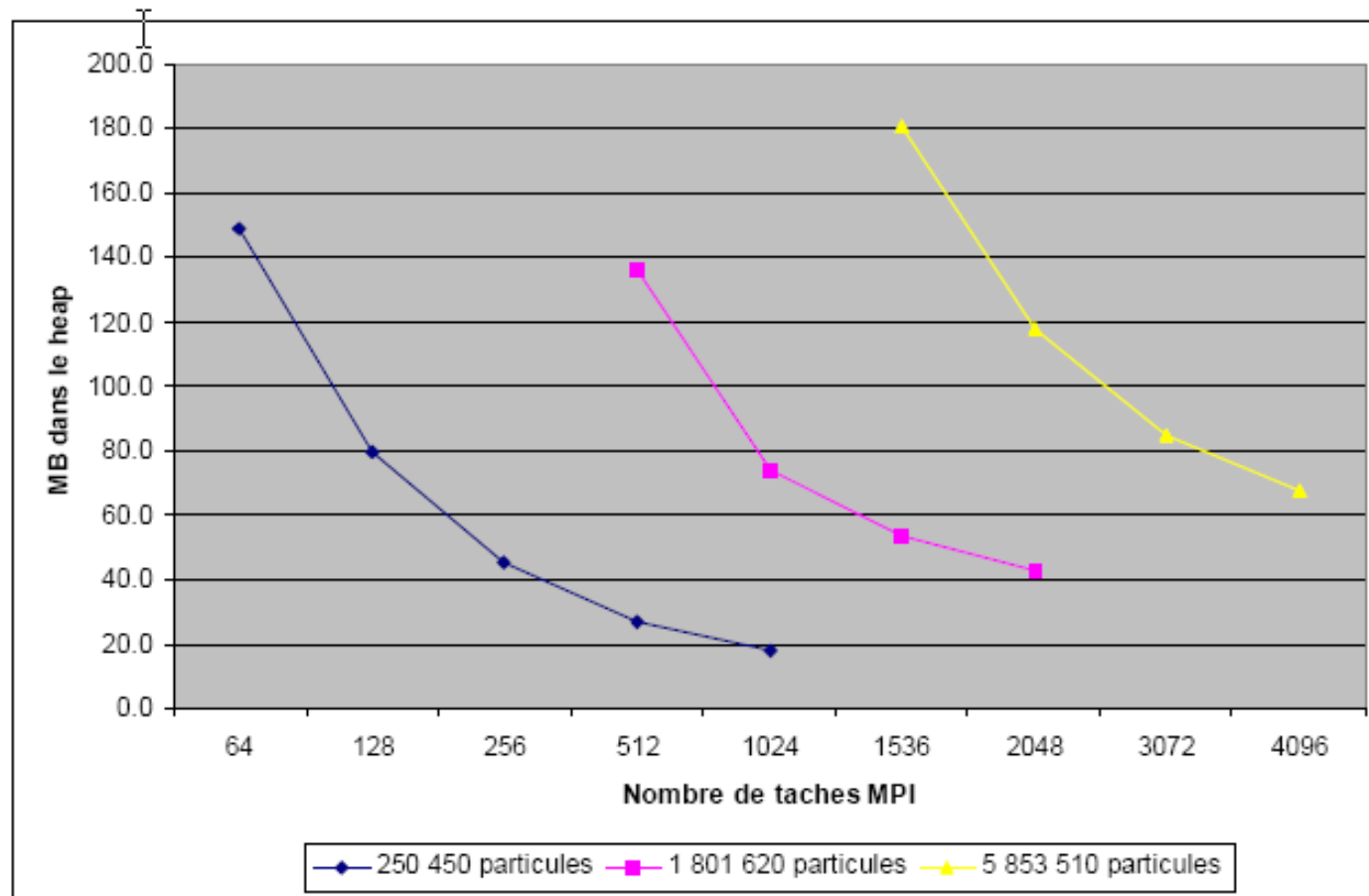


Performances on Blue Gene for 250k particles

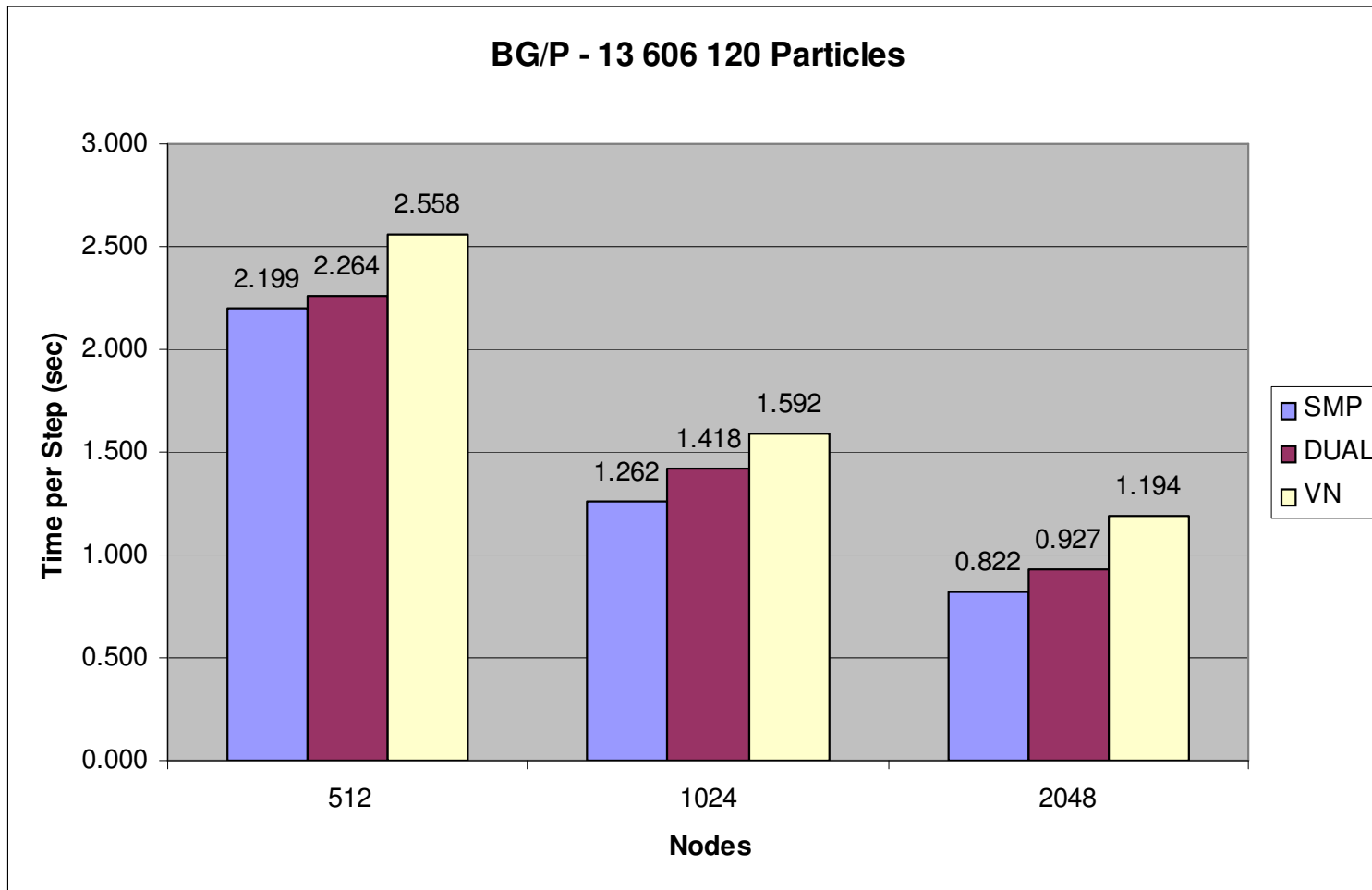
Scalability comparisons with version January 2008

(improved since)

# Memory footprint per process



## Mixed Mode Implementation on Blue Gene/P



# Main Features of NEPTUNE\_CFD

## The CFD Scale of the NEPTUNE Platform

### 3D and local two-phase flow analysis

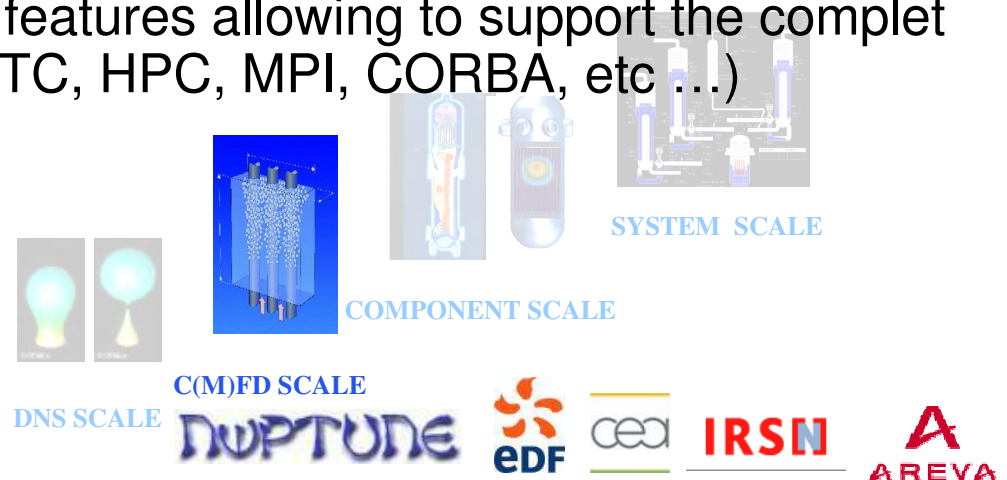
- Generalized multi-field model
- Turbulence, IAT, non condensable gases

### Development team: EDF and CEA

- Development, validation, maintenance, installation, training, hot-line

### IBM Involvement

- Support Blue Gene porting and tuning
- Provide the necessary BG features allowing to support the complete production environment (HTC, HPC, MPI, CORBA, etc ...)

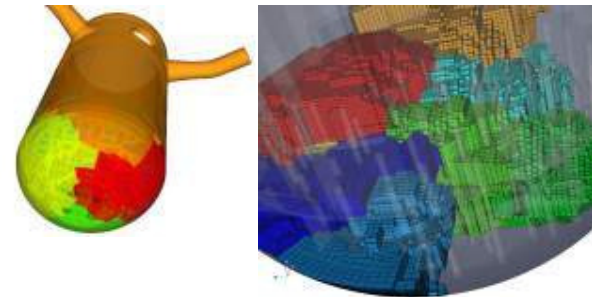


# Main Features

## Parallel Software Architecture

### Distributed memory approach

- SIMD technique
- Domain partitioning (METIS)
- Message passing (MPI)



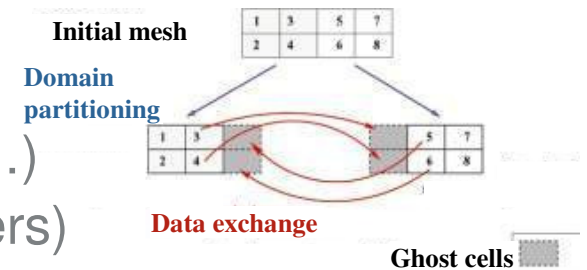
### From 2 Proc-PCs to 8000-Proc Blue-Gen...

- ...through local or remote clusters



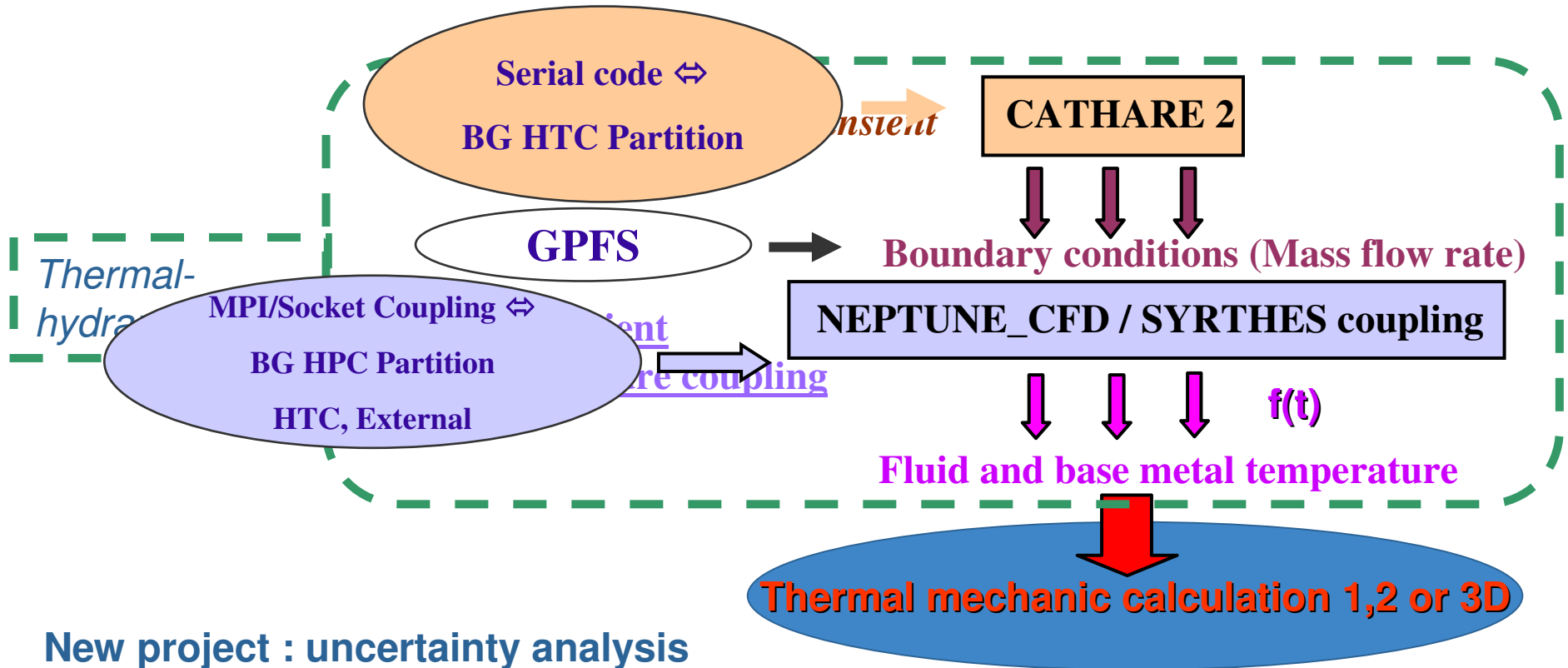
### Limited impact for developers

- I/O operations
- Global operations (dot-products, min, max...)
- Linear system solves (sparse iterative solvers)
- Flux computation using neighbouring cells



# Context : multiple-fied integrity studies & Blue Gene Requirements

Standard methodology involves chained and coupled computations



## New project : uncertainty analysis

- a Monte-Carlo method is used requiring  $n$  thermal-hydraulics steps
- The CPU consuming part is the NEPTUNE\_CFD / SYRTHES part (3D CFD)

NEPTUNE\_CFD

## **CATHARE (Serial) simulation and IBM Blue Gene HTC capability**

- **Objective: run a series of CATHARE serial jobs one Blue Gene racks**
  - ▶ **3-6 days of simulation time and ~ 10Tbytes of data**
    - Serial execution time varies from 1 minute to 6 days
    - Large number of runs > 5000
  - ▶ **Communication with Neptune through files (GPFS)**
  
- **IBM is in charge to develop CATHARE interfaces for Blue Gene HTC environment**

## IBM Blue Gene HTC partition

**Blue Gene 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
- High Throughput Computing (HTC) Model
  - Large number of independent tasks
  - Programming model: non-MPI
  - Support all the modes on BGP ( CMP, DUAL, VN)
  - Keep running with node failures

**Blue Gene/P HTC brings a lot of improvements compared to BG/L**

## HTC on Blue Gene\L (very simple feature)

- **HTC Launcher – resident on each BG node**
  1. Listens on socket for work-requests from HTC Scheduler
  2. Performs exec command to launch the job
  3. Restarts the launcher!
- **HTC Scheduler**
  - ▶ Transfers work-request to HTC Launcher collective
- **Users must implement the HTC launcher and Scheduler**
- **No Socket Client support ⇔ no direct communication in a HTC partition**
- **Limit performance on Virtual mode, no socket multiplexing, no stderr/stdin**

## HTC on Blue Gene\P (fully integrated)

- Simple command “*submit*” (similar to mpirun for HPC partition) with stdin/stdout support
- Fully Socket implementation ⇔ possible TCP communications inside and outside HTC partition (HTC-HTC, HTC-HPC, HTC-Frontend)
- Socket multiplexing, one socket per IO node
- “Simple Scheduler” on IBM AlphaWork
- CONDOR Collaboration (not yet fully integrated)

## IBM HTC SIMPLE Scheduler Overview

- **SIMPLE scheduler provides features not available with “submit” interface**
  - ▶ Provides queuing of jobs until compute resources are available
  - ▶ Tracks failed compute nodes
  - ▶ “submit” interface is intended for usage by job schedulers ... not end users directly
- **simple\_sched daemon**
  - ▶ Runs on service node (SN) or front-end node (FEN)
  - ▶ Accepts connections from startd and client programs
- **startd daemons**
  - ▶ Run on submit nodes (FEN typically)
  - ▶ Connects to simple\_sched, gets jobs and executes submit
- **Client programs**
  - ▶ qsub – Submits job to run
  - ▶ qdel – Deletes job submitted by qsub
  - ▶ qstat – Gets status of submitted job
  - ▶ qcmd – Admin commands



## Very easy to integrate with LoadLeveler

- ▶ Idea is to use LL to create partition but not to BOOT partition
  - ▶ End user can use a static partition or have LL dynamically create a partition
  
- ▶ LoadLeveler Job Command File (JCF) for HTC jobs requires the following:
  - ▶ **htcpartition** executable:
    - Utility program shipped in Blue Gene
    - Responsible for booting and freeing HTC partition from a FEN
  - ▶ **run\_simple\_sched\_jobs** executable
    - Provides “personal” instance of SIMPLE job scheduler and startd
    - Executes commands either specified in command files or read from stdin
    - Creates a cfg file that can be used to submit jobs externally to the cmd files or stdin
    - Exits when the commands have all finished (or can specify “keep running”)

## Allowing multiple users to run jobs on same partition with glide-in

- The **htcpartition** keyword `--userlist <list|*ALL>` indicates who can run jobs on the partition
  - The JCF must have **htcpartition** specified with those “other” users or `*ALL` for anyone (A)
- “Other” users must know the configuration of the “personal” instance of SIMPLE to run jobs
  - Easiest way is to share the configuration file name from the JCF (e.g. `my_cfg.cfg`) (B)
  - External users then use **qcmd** pointing to active `simple_sched` (e.g. `qcmd --config my_cfg.cfg`) (C)

### “Other” Users (C)

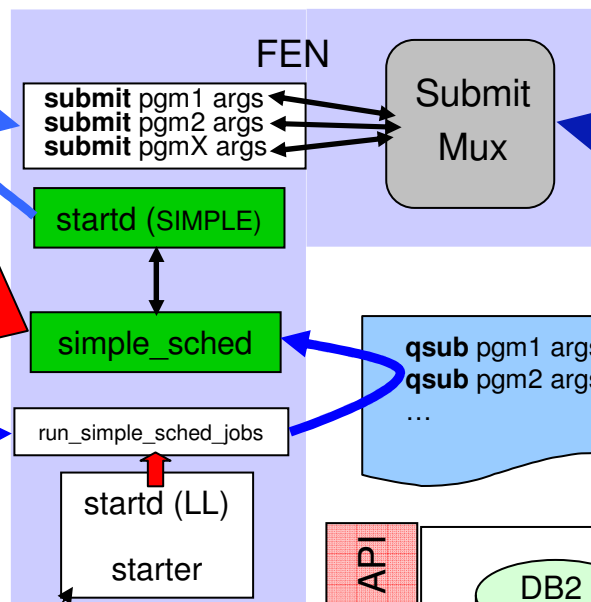
`qcmd --config my_cfg.cfg`

```
pgmX args
pgmY args
pgmZ args
...
```

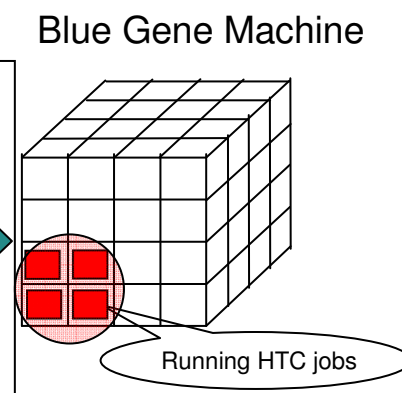
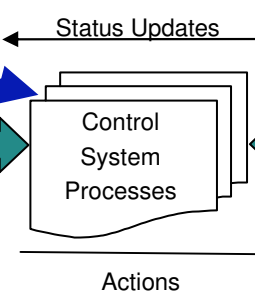
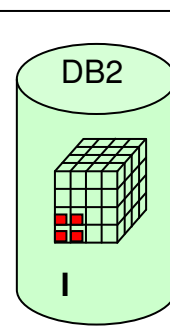
*cmds.run file*

```
pgm1 args
pgm2 args
pgm3 args
...
```

Service node  
Central Manager

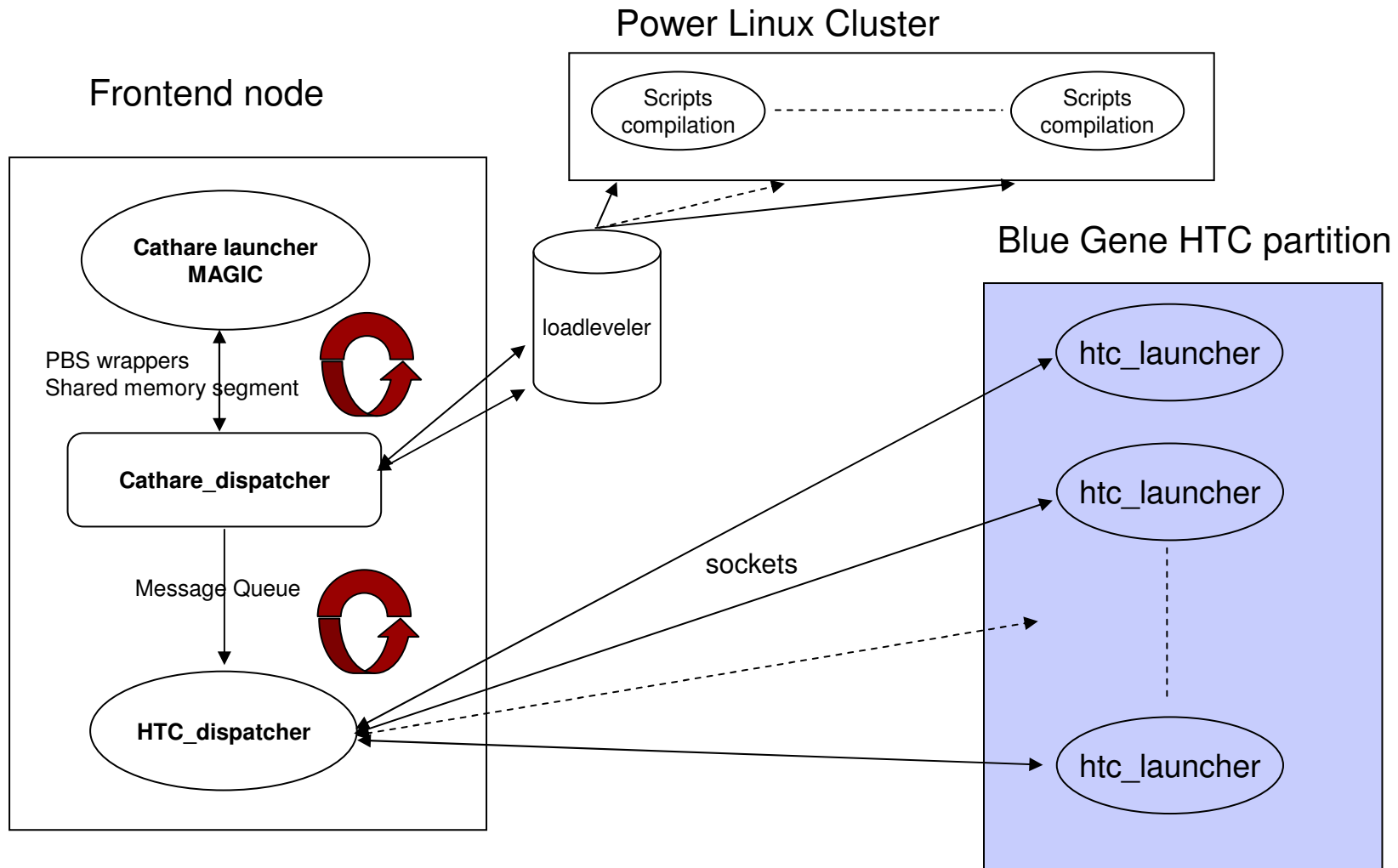


Control System API



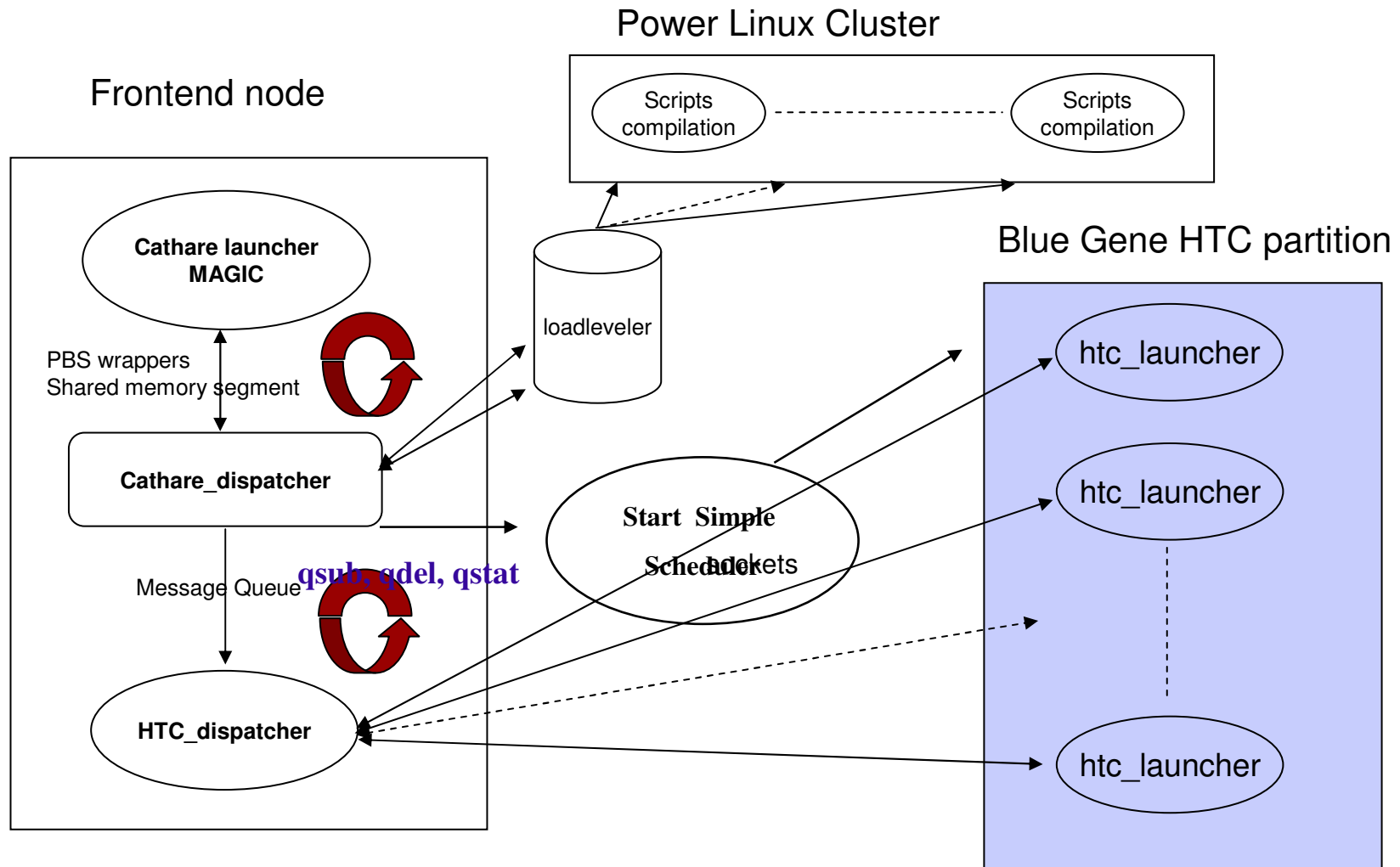
```
# @ job_type = bluegene
# @ queue
htcpartition --boot --mode SMP
  --userlist ELVIS,FRANK,DEAN (A)
  (B) --configfile my_cfg.cfg
run_simple_sched_jobs
  (B) -config my_cfg.cfg cmds_run
```

# How does CATHARE work on Blue Gene/L (job submission only)



## How does CATHARE work on Blue Gene/P

Replace HTC\_dispatcher and htc\_launcher user routines by IBM simple scheduler (qsub command), with more flexibility and RAS features + IBM support



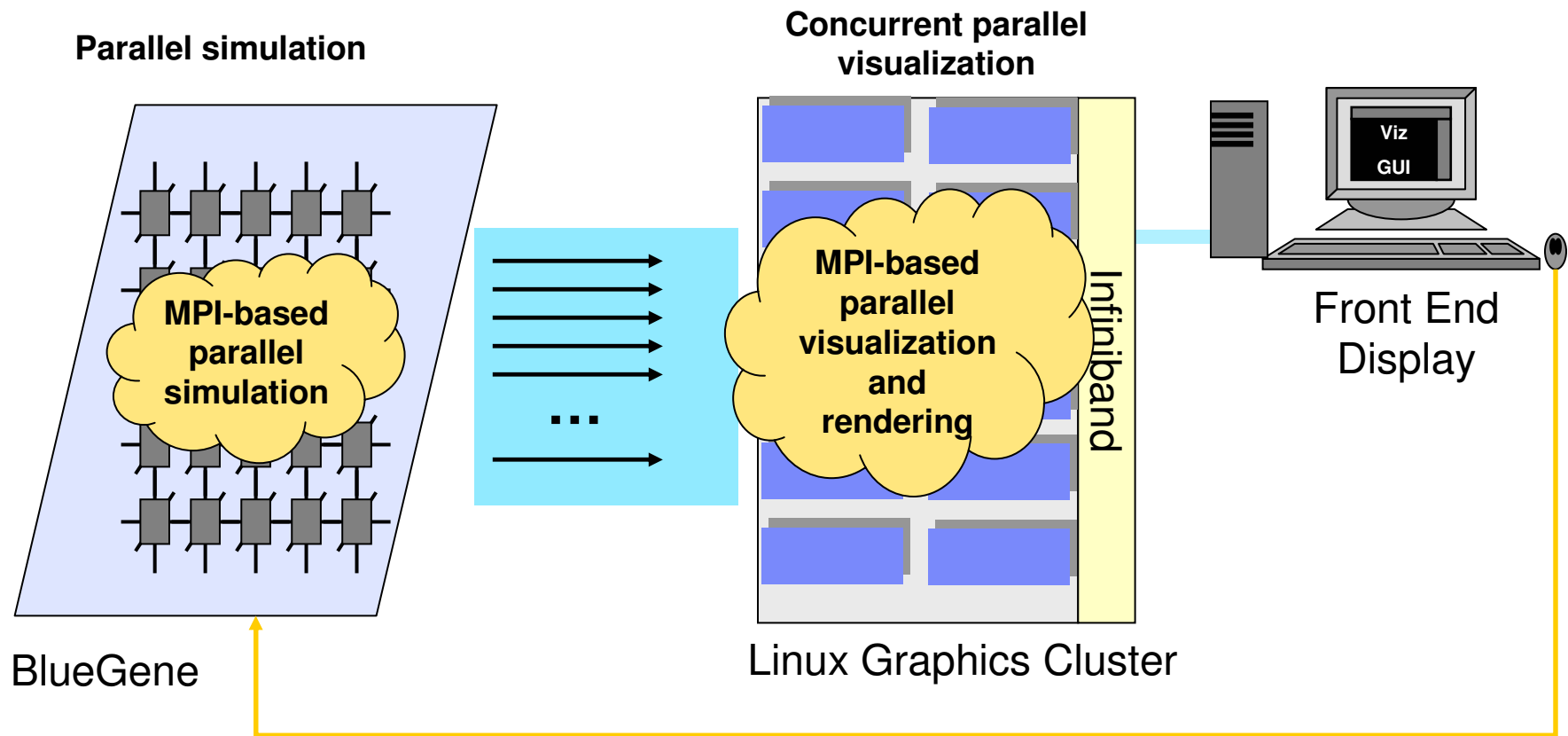
***Thank for your attention***

***QUESTION ?***

➤ BACKUP

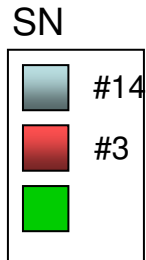
# BlueGene as a Computational Accelerator Linux Cluster as a Graphics Accelerator

**2 types of Connections : GPFS (files) or Sockets (no file)**



# Blue Gene/L

|         |     |      |
|---------|-----|------|
| N7 J117 | #23 | # 24 |
| N6 J115 | #21 | #22  |
| N5 J113 | #19 | #20  |
| N4 J111 | #17 | #18  |
| S0 J109 | # 1 | #2   |
| N3 J108 | #15 | #16  |



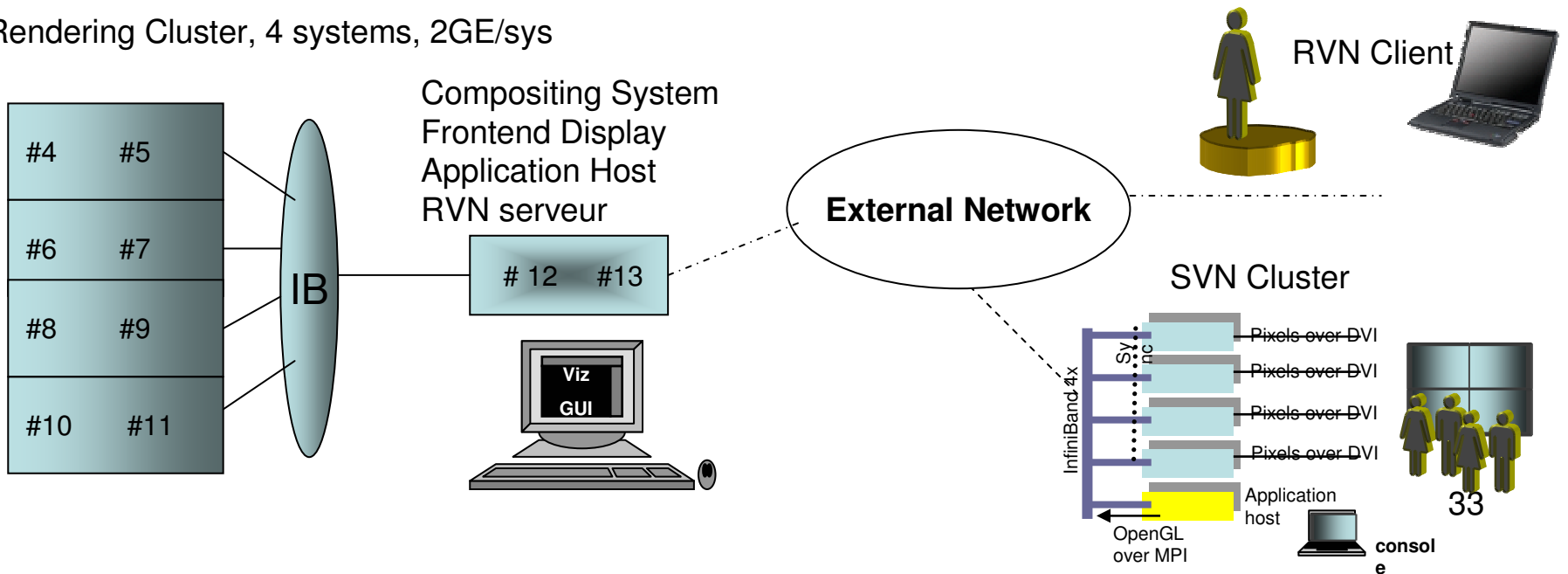
# Blue Gene/L and Visualization configuration Demo

Internal Network: 24 ports - Switch configuration

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

## DCV / SPVN Cluster

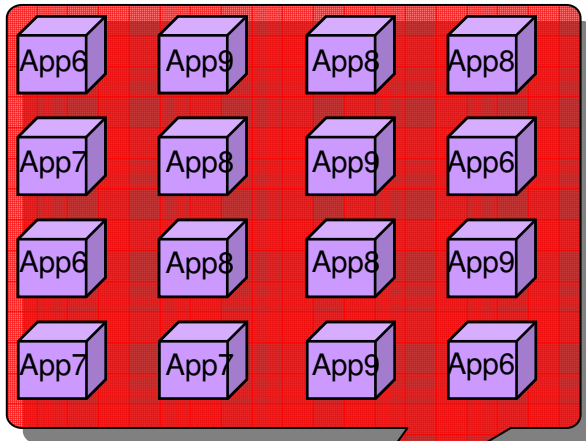
Rendering Cluster, 4 systems, 2GE/sys



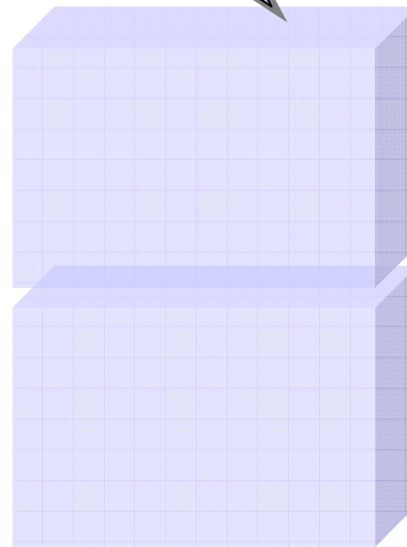
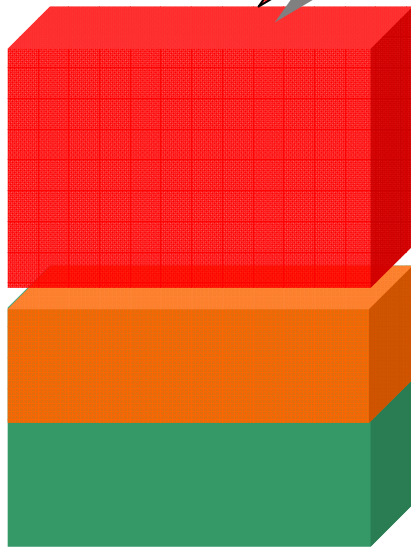
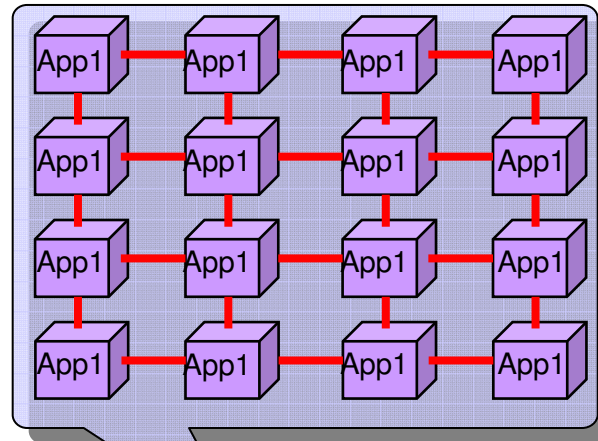
# Partitioning example: mixing HTC and HPC







**HTC**

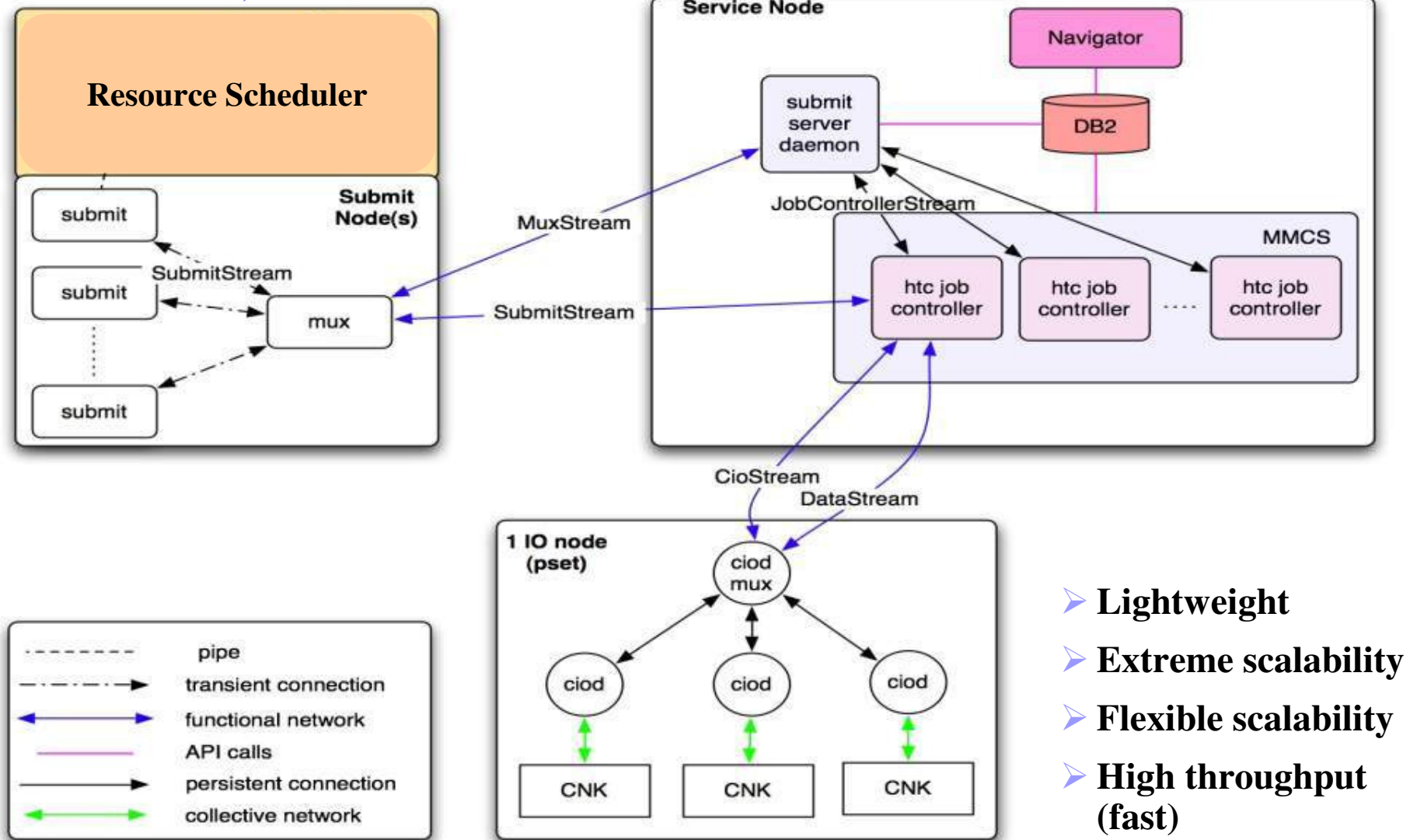


**HPC**



-  HPC – 1,024 nodes
-  HTC Virtual Node Mode – 512 nodes
-  HTC Dual Mode - 256 nodes
-  HTC SMP Mode - 256 nodes

## How does HTC work on BGP (fully integrated through submit command)



- **Lightweight**
- **Extreme scalability**
- **Flexible scalability**
- **High throughput (fast)**

## SIMPLE Scheduler Available through IBM alphaWorks

