How to Use HPC Resources Efficiently by a Message Oriented Framework

www.hp-see.eu

HP-SEE

High-Performance Computing Infrastructure for South East Europe's Research Communities

E. Atanassov, T. Gurov, A. Karaivanova
Institute of Information and Communication Technologies
Bulgarian Academy of Science
(emanouil, gurov, anet)@parallel.bas.bg

Supported by Grant #DO 02-146
with NSF of Bulgaria

The HP-SEE initiative is co-funded by the European Commission under the FP7 Research Infrastructures contract no. 261499

OUTLINE

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- Motivation
- Bulgarian HPC resources
- Interconnection of the existing HPC platforms
- Case study – MC modelling of semiconductor devices
- Message oriented framework
- Conclusions and future work

Motivation

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

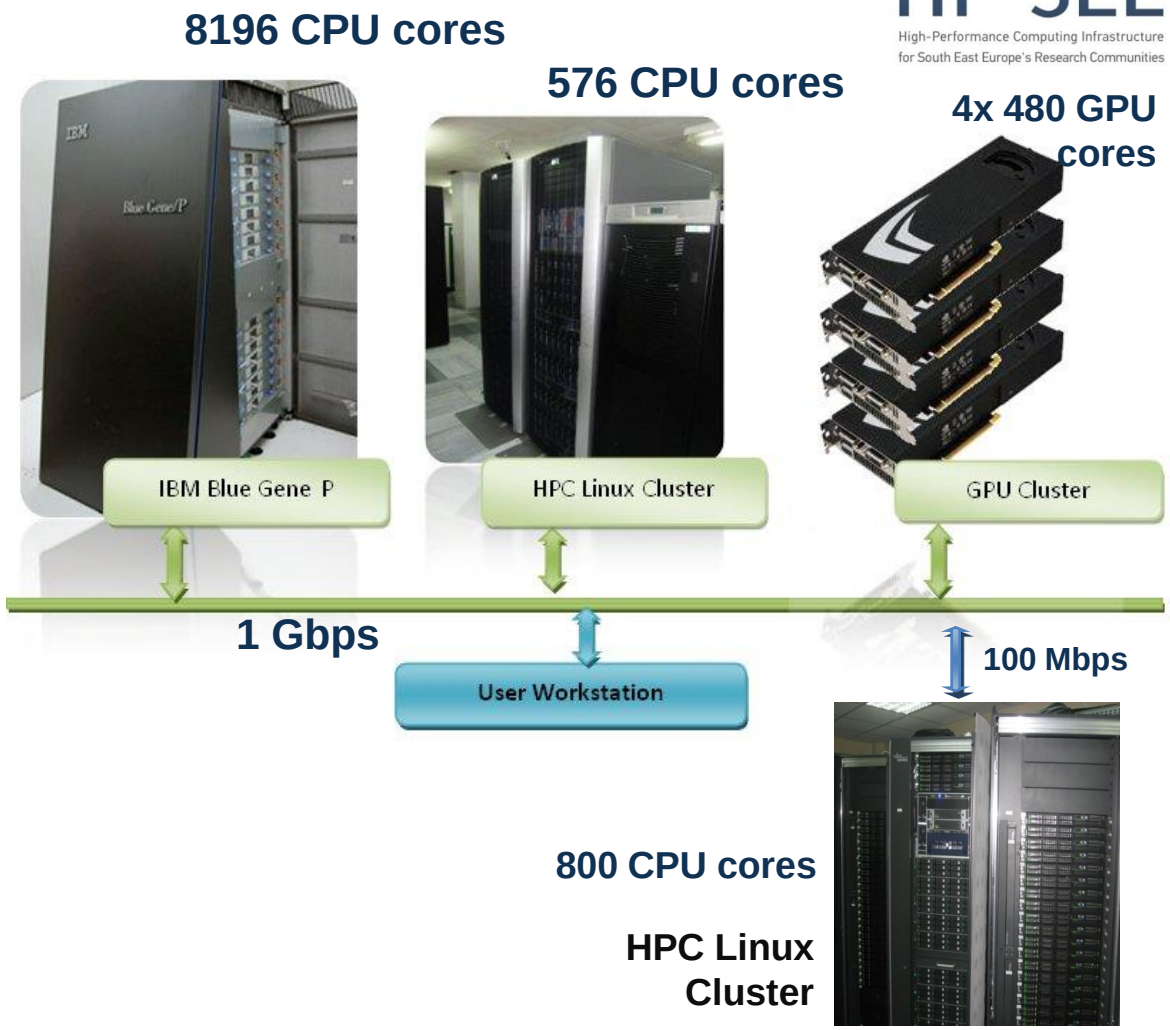
- Significant imbalance between computational power and I/O bandwidth of current leadership-class machines
- Porting an application to a new platform can be challenging
- A researcher may have access to several different HPC resources at the same time

  Here we will go through some points that should be taken into consideration when porting applications to HPC recourses in Bulgaria. We suggest Message Oriented Framework with Low Overhead for efficient use of the HPC recourses available in Bulgaria

Bulgarian HPC Resources

- The biggest HPC resource for research in Bulgaria is the supercomputer - BlueGene/P at EA "ECNIS"
  - - vendor: IBM
- Two HPC clusters with Intel CPUs and Infiniband interconnection at IICT-BAS and IOCCP-BAS
  - - vendors: HP and Fujitsu
- In addition servers equipped with powerful GPU are available for applications that can take advantage of them.
- 1 Gb/s Ethernet fiber optics links

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities



BG Blue Gene/P in Sofia

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- IBM Blue Gene/P –two racks, 2048 *PowerPC 450 processors (32 bits, 850 MHz), a total of 8192 cores;*
- A total of 4 TB random access memory;
- **16 I/O nodes** currently connected via fiber optics to a **10 Gb/s** Ethernet switch;
- Theoretical peak performance: Rpeak= **27.85 Tflops;**
- **Energy efficiency: 371.67 MFlops/W**
- 1 Gb/s Ethernet fiber optics link to Bulgarian NREN's Point-of-Presence at the IICT-BAS
- Operating System for front-end node: SUSE Linux Enterprise Server 10 (SLES 10), Service Pack 1 (BG/P)
- The Compute Nodes run OS Compute Node Kernel (CNK)
- 2 file servers, 12 TB storage

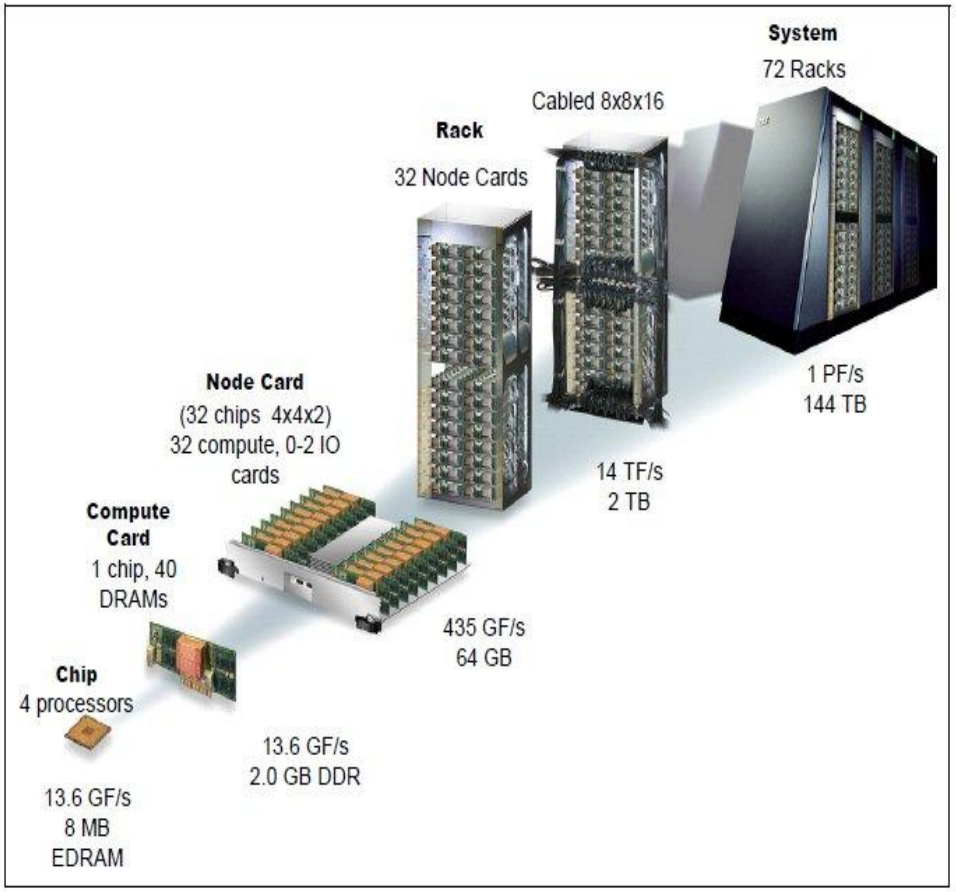


Figure 1-2 Blue Gene/P packaging

Networks on Blue Gene/P

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- On the BG/P five networks are used for various tasks.
- #1/ **Three-dimensional torus**: for general-purpose, point-to-point message passing. A fully connected torus is only available when using jobs which occupy one or two complete midplanes, i.e. half or all of the system. When using smaller partitions of the machine, the network's topology is degraded to a 2 or 3D mesh.
- #2/ **Global collective network** is a high-bandwidth, tree based network that is used for collective communication operations, such as broadcasts and reductions, and to move process and application data from the I/O nodes to the compute nodes.
- #3/ **Global interrupt network**
- #4/ **10 Gigabit Ethernet network** uses optical cabling to interconnect all the I/O nodes and the storage infrastructure.
- #5/ **Control:** The control network consists of a JTAG interface with direct access to shared SRAM in every compute and I/O node. It is used for boot, monitoring, and diagnostics.

Existing infrastructure – BG HPC Cluster at IICT-BAS

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- HP Cluster Platform Express 7000 enclosures with 36 blades BL 280c with dual Intel Xeon X5560 @ 2.8Ghz (total **576** cores), 24 GB RAM per blade
- 8 controlling nodes HP DL 380 G6 with dual Intel X5560 @ 2.8 Ghz, 32 GB RAM
- Non-blocking DDR Interconnection via Voltaire Grid director 2004
- Two SAN switches for redundant access
- MSA2312fc with **96 TB storage**, Lustre filesystem.
- More than 92% efficiency on LINPACK (>3 TFlops, peak performance 3.2TFlops)
- SL 5, gLite, torque+maui, eucalyptus walrus
- **Connected to it – extension cluster with 4 GPU cards NVIDIA GTX 295**

(4x2x240 GPU cores =1920 cores, 576 MHz, 1.8 GB per card)

Interconnection of the existing HPC platforms

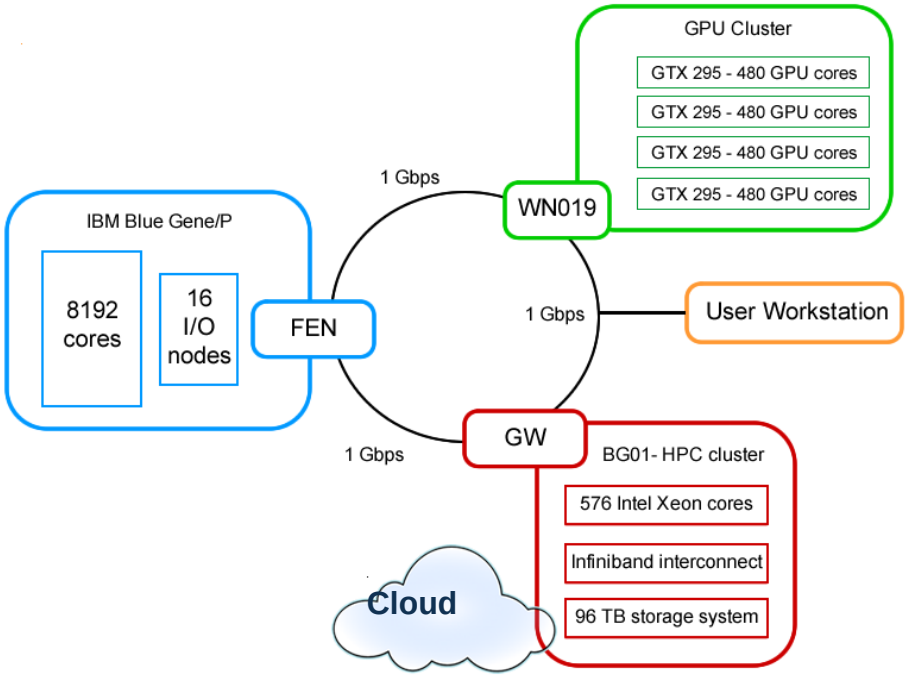
HP-SEE

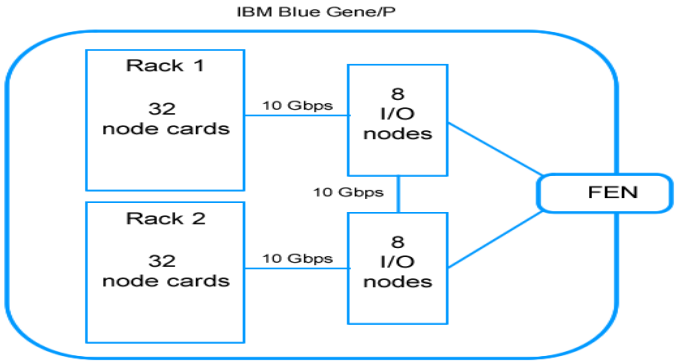
rastructure
ommunities

The diagram shows the interconnection of existing HPC platforms in Bulgaria:

- Front-end Nodes (fen, GW, WN019) provide access to the users to develop, build and submit
- Compute Nodes run applications
- I/O Nodes provides access to external devices. All I/O requests are routed through these nodes. (BG/P))
- All I/O calls in the applications are forwarded to the I/O nodes.
- High bandwidth networks for end-user applications use, i.e. point to point & collective communications.

In order to facilitate the coordinated use of all these resources where each resource is used for the parts of the application where it is most efficient, we have developed a framework that allows the researcher to interconnect resources of the above types with minimal overhead





Case Study: Ultra-fast Transport in sEmiconductors

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- Application area:
  - Simulation of Electron Transport (SET) is developed for solving various computationally intensive problems which describe ultrafast carrier transport in semiconductors.
- Expected results and their consequences
  - studies memory and quantum effects during the relaxation process due to electron-phonon interaction in semiconductors; present version explores electron kinetics in GaAs nano-wires.
  - Studying the quantum effects that occur at nanometer and femtosecond scale have important scientific results - novel advanced methods, investigation of novel physical phenomena

Quantum-kinetic equation (inhomogeneous case)

HP-SEE
High-Performance Computing Infrastructure
arch Communities

The integral form of the equation:

$$f_w(z,k_z,t) = f_{w,0}(z - \frac{\hbar k_z}{m}t, k_z) +$$

$$+\int_0^t dt'' \int_{t''}^t dt' \int_G d^3\mathbf{k}' \{K_1(k_z,\mathbf{k}',t',t'') f_w\left(z + h(k_z,q_z',t,t',t''), k_z', t''\right)\}$$

$$+\int_0^t dt'' \int_{t''}^t dt' \int_G d^3\mathbf{k}' \{K_2(k_z,\mathbf{k}',t',t'') f_w\left(z + h(k_z,q_z',t,t',t''), k_z, t''\right)\}$$

$$h(k_z,q_z',t,t',t'') = -\frac{\hbar k_z}{m}(t-t'') + \frac{\hbar q_z'}{2m}(t'-t'')$$

Kernels:

$$K_1(k_z,\mathbf{k}',t',t'') = S(k_z',k_z,t',t'',\mathbf{q}_\perp') = -K_2(\mathbf{k}',k_z,t',t'')$$

$$S(k_z',k_z,t',t'',\mathbf{q}_\perp') = \frac{2V}{(2\pi)^3}|G(\mathbf{q}_\perp')\mathcal{F}(\mathbf{q}_\perp',k_z-k_z')|^2 \times$$

$$\left[(n(\mathbf{q}')+1)\cos\left(\frac{\epsilon(k_z)-\epsilon(k_z')+\hbar\omega_{\mathbf{q}'}}{\hbar}(t'-t'')\right)\right.$$

$$\left. +n(\mathbf{q}')\cos\left(\frac{\epsilon(k_z)-\epsilon(k_z')-\hbar\omega_{\mathbf{q}'}}{\hbar}(t'-t'')\right)\right]$$

Quantum-kinetic equation (cont.)

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

Bose function: $n_{\mathbf{q}'} = 1/(\exp(\hbar\omega_{\mathbf{q}'}/\mathcal{K}T) - 1)$

The phonon energy (*ħω*) depends on : $\mathbf{q}' = \mathbf{q}'_\perp + q'_z = \mathbf{q}'_\perp + (k_z - k'_z)$

Electron energy: $\varepsilon(k_z) = (\hbar^2 k_z^2)/2m$

The electron-phonon coupling constant according to Fröhlich polar optical interaction:

$$\mathcal{F}(\mathbf{q}'_\perp, k_z - k'_z) = -\left[\frac{2\pi e^2 \omega_{\mathbf{q}'}}{\hbar V}\left(\frac{1}{\varepsilon_\infty} - \frac{1}{\varepsilon_s}\right)\frac{1}{(\mathbf{q}')^2}\right]^{\frac{1}{2}}$$

The Fourier transform of the square of the ground state wave function:

$$G(\mathbf{q}'_\perp) = \int d\mathbf{r}_\perp e^{i\mathbf{q}'_\perp \mathbf{r}_\perp} |\Psi(\mathbf{r}_\perp)|^2$$

$$|G(\mathbf{q}'_\perp)|^2 = |G(q'_x)G(q'_y)|^2 =$$

$$\left(\frac{4\pi^2}{q'_x a\left((q'_x a)^2 - 4\pi^2\right)}\right)^2 4\sin^2(aq'_x/2)\left(\frac{4\pi^2}{q'_y a\left((q'_y a)^2 - 4\pi^2\right)}\right)^2 4\sin^2(aq'_y/2)$$

Monte Carlo method

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

$$J_g(f) \equiv (g,f) = \int_0^{\mathcal{T}} \int_D g(z,k_z,t) f_w(z,k_z,t) dz dk_z dt$$

$$(z,k_z) \in D = (-Q_1,Q_1) \times (-Q_2,Q_2), \quad t \in (0,\mathcal{T})$$

$$(i) \quad g(z,k_z,t) = \delta(z-z_0)\delta(k_z-k_{z,0})\delta(t-t_0)$$

$$(ii) \quad g(z,k_z,t) = \frac{1}{2\pi}\delta(k_z-k_{z,0})\delta(t-t_0)$$

$$(iii) \quad g(z,k_z,t) = \frac{1}{2\pi}\delta(z-z_0)\delta(t-t_0)$$

Backward time evolution of the numerical trajectories

Wigner function:

$f_w(z,k_z,t)$

Energy (or momentum) distribution:

$$f(k_z,t) = \int \frac{dz}{2\pi} f_w(z,k_z,t)$$

Density distribution:

$$n(z,t) = \int \frac{dk_z}{2\pi} f_w(z,k_z,t)$$

Monte Carlo Method (cont.)

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

Biased MC estimator:

Weights: $$W_j^\alpha = W_{j-1}^\alpha \frac{K_\alpha(k_{zj-1}, \mathbf{k}_j, t'_j, t_j)}{p_\alpha p_{tr}(\mathbf{k}_{j-1}, \mathbf{k}_j, t'_j, t_j)},\ W_0^\alpha = W_0 = 1,\ \ \alpha = 1,2,\ \ j = 1, \ldots, s$$

$$(k_{zj}, t'_j, t_j) \in (-Q_2, Q_2) \times (t_j, t_{j-1}) \times (0, t_{j-1})$$

The Markov chain:

$$(k_{z0}, t_0) \to (k_{z1}, t'_1, t_1) \to \ldots \to (k_{zj}, t'_j, t_j) \to \ldots \to (k_{zs}, , t'_s, t_s),\ j = 1, 2, \ldots, s$$

$$(z, k_{z0}, t_0): \quad p_{in}(z, k_z, t) = g(z, , k_z, t)$$

Initial density function

$$p_{tr}(\mathbf{k}, \mathbf{k}', t', t'') = p(\mathbf{k}'/\mathbf{k}) p(t, t', t'')$$

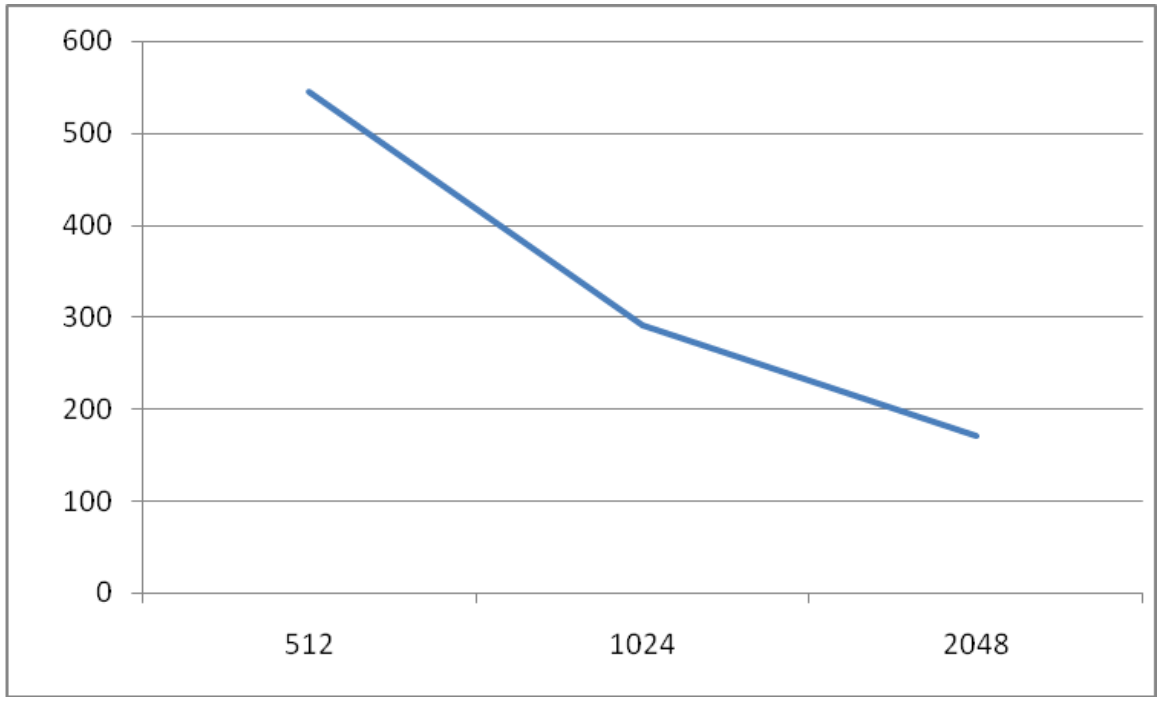
Transition density function:

$$1/N \sum_{i=1}^{N} (\xi_s[J_g(f)])_i \to J_g(f)$$

SET scalability on Blue Gene/P

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- SET was tested on Bulgarian BlueGene/P and showed the following scalability results on 512, 1024 and 2048 cores. The test case uses 10 millions of trajectories to simulation 180 femtosecond evolution.



Case study: ultrafast transport in semiconductors

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- The application requires accumulation of billions of trajectories
- Improvements in variance and execution time can be achieved with low-discrepancy sequences (quasirandom numbers).
- The use of quasirandom numbers requires a robust and flexible implementation, since it is not feasible to ignore failures and missing results of some trajectories, unlike in Monte Carlo.
- GPU resources are efficient in computations using the low-discrepancy sequences of Sobol, Halton, etc.
- Our idea is to use exchange of messages for distributing the work between heterogeneous HPC resources independent of underlying middleware

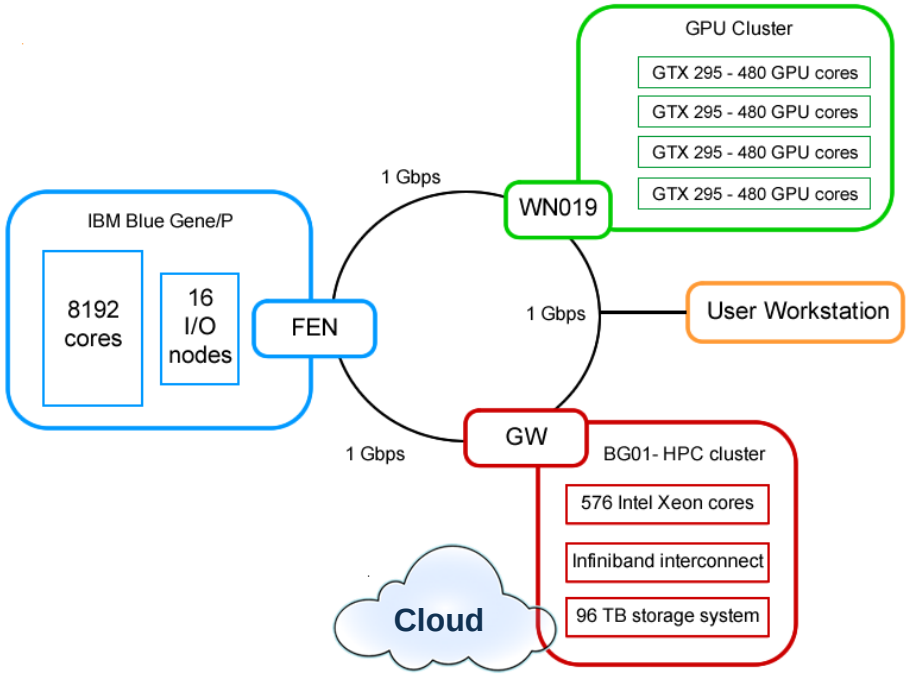
Message oriented framework

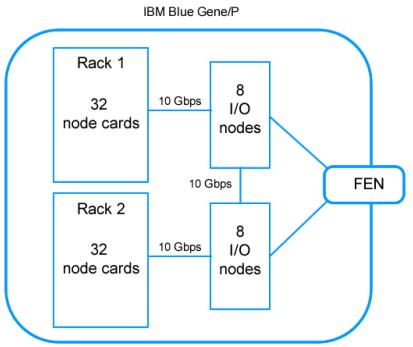
HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- The idea is to create a framework that is sufficiently generic to be used by a wide range of applications but is efficient for the case of semiconductor modelling with MC and QMC.
- Two patterns for exchange of information:
  - Short messages, encoding parameters and results (RPC type)
  - Large amount of data, organized as file and using permanent or temporary disk storage

Message oriented framework

- Most of HPC computing resources are hidden behind firewalls, communications with external nodes must be forwarded through the gateway nodes, where user-level server program can be started by the user.
- User starts the server programs at the gateways and launches computational batch jobs.





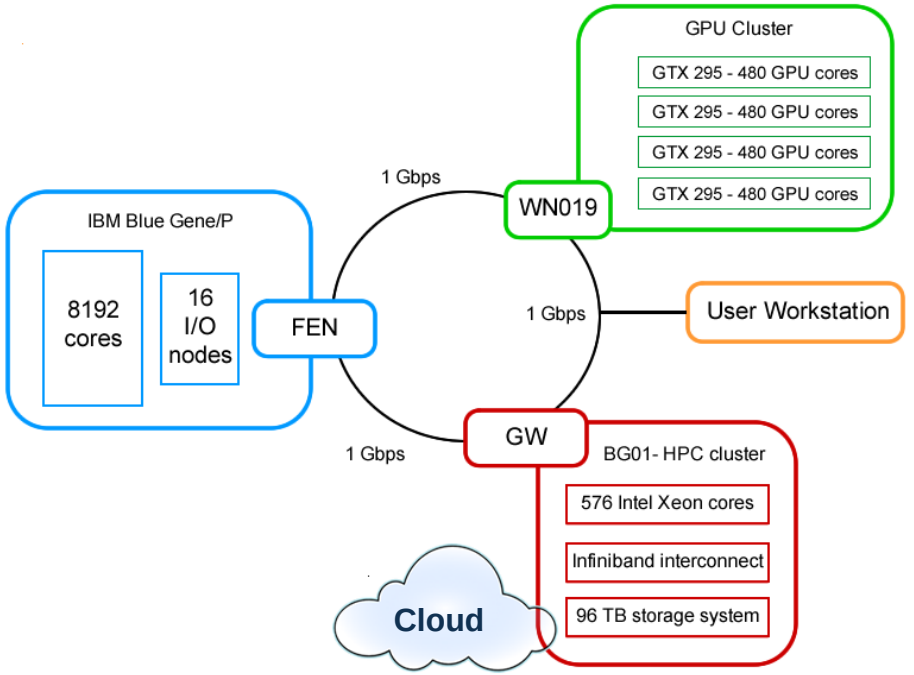
Message oriented framework

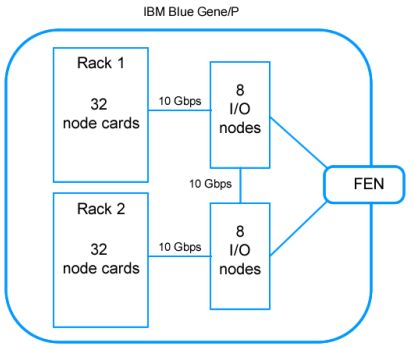
- The computational jobs connect with the server and follow the pattern:
  - Request work
  - Return result

  until receiving marker for end of computation.

  The server performs match-making of requests, for example some kinds of requests may be served only by particular resources.

  One server (wn019) is master, the others forward to it.





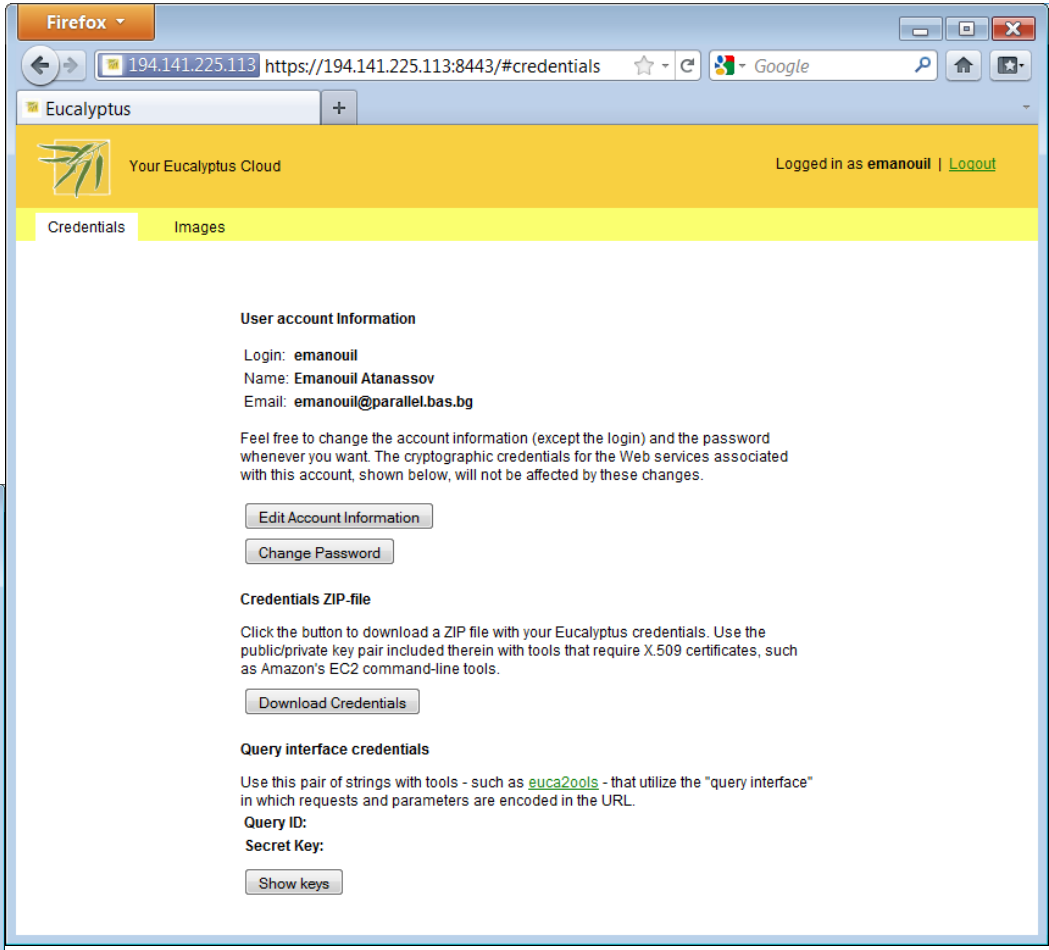
Using cloud storage

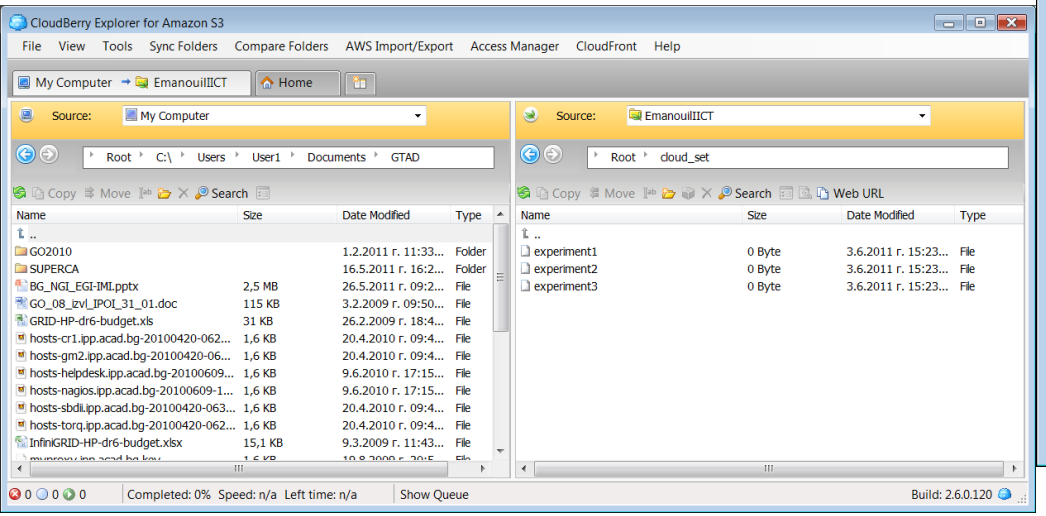
HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

Users register at web portal and obtain access to cloud storage at IICT-BAS

Access via windows or linux app or from command line

(s3curl was ported to work on Blue Gene/P)





Conclusions and future work

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- Message oriented frameworks overcome some deployment limitations like lack of common Grid middleware installed
- Access to cloud storage provides simple security model (signed http requests) which also offers easier deployment
- Robustness of the framework should be improved with regards to dealing with failures and timeouts.