Meta-computations on the CLUSTERIX Grid

Roman Wyrzykowski[1], Norbert Meyer[2], Tomasz Olas[1], Konrad Karczewski[1], Bogdan Ludwiczak[2], Cezary Czaplewski[3], and Stanislaw Oldziej[3]

[1] Dept. of Comp. & Information Science, Czestochowa Univ. of Technology, Poland
[roman, olas, xeno]@icis.pcz.pl
[2] Poznan Supercomputing and Networking Center, Poland
[meyer, bogdanl]@man.poznan.pl
[3] Department of Chemistry, University of Gdansk, Poland
czarek@sun1.chem.univ.gda.pl

Abstract. In the first part of the paper, we present the concept and implementation of the National Cluster of Linux System (CLUSTERIX) – a truly distributed national computing infrastructure with with 12 sites (local Linux clusters with 64-bit architecture) located accross Poland. The second part presents our experience in adaptation of selected scientific applications to the cross-site execution as meta-applications, using the MPICH-G2 environment. The performance results of experiments confirm that CLUSTERIX can be an efficient platform for running meta-aplications. However, harnessing its computing power needs to take into account the hierachical architecture of the infrastructure, and its heterogeneity.

1 CLUSTERIX Grid Project

The main objective of the CLUSTERIX national grid project [1, 8] is to develop mechanisms and tools that allow for deployment of a production grid environment, with the backbone consisting of dedicated, local Linux clusters with 64-bit machines. Local clusters are placed across Poland, and connected by the Polish Optical Network PIONIER. The resulting infrastructure should be tested on a set of pilot applications developed as a part of the project. It is implemented by 12 Polish supercomputing centers and metropolitan area networks, with Czestochowa University of Technology as a coordinator.

CLUSTERIX infrastructure is a distributed PC-cluster (or meta-cluster) with 12 local Linux clusters in the backbone, located in independendent centers across Poland. They are interconnected via dedicated 1 Gb/s channels provided by PIONIER. At this moment, the CLUSTERIX backbone includes 254 Intel Itanium2 processors (1.3 GHz, 3 MB cache) in 127 two-way SMP nodes. Two Gbits-per-second VLANs are used to improve management of network traffic in local clusters.

Selected 32-bit machines are dedicated to management of local clusters and the entire infrastructure. While user tasks are allowed to be executed only on computational nodes, each local cluster is equipped with an access node where the Globus Toolkit [3] and local batch system are running. An important element of the CLUSTERIX backbone is the Data Storage System managed by

the CLUSTERIX Data Management System. Another example of middleware developed in the project is the Virtual Users' Account System which allows for efficient management of users' acounts and virtual organizations.

The CLUSTERIX middleware is developed as Open Source, and is based on the Globus Toolkit 2.4 and Web Services, with Globus 2.4 available in the Globus 3.2 distribution. The use of Web Services makes the created software easier to reuse, and allows for interoperability with other grid systems on the service level. The important feature of this middleware is abilty to manage infrastructure with dynamic changing configuration. In particular, new clusters may be attached to the CLUSTERIX backbone dynamically. The connection of new clusters to the backbone opens possibilities to access a shared environment with the extraordinary computational power. For example, an experimental installation with 802 Itanium2 CPUs offering a peak performance of about 4,5 TFLOPs has been created.

2 Meta-applications in CLUSTERIX

According to the project goals, CLUSTERIX is used both for running HTC applications, as well as large-scale distributed applications that require parallel use of one or more local clusters. Selected end-user applications have been deployed in the CLUSTERIX grid. It is clear that delivering applications able to use distributed resources efficiently will in the end decide on success or failure of computational grids.

These applications fall into two different categories:

- Applications dedicated to run on a single local cluster; this restriction is related only to a single instance of an application, while different instances can be executed on different local clusters.
- Meta-applications intended to be executed on more than one local cluster (cross-site execution); they will take into account the heterogeneity of the meta-cluster.

2.1 Using MPICH-G2

In the CLUSTERIX project, the MPICH-G2 middleware [5] based on the Globus Toolkit is used as a grid-enabled implementation of the MPI standard. It allows for running multilevel parallel applications across many sites [4]. MPICH-G2 extends the MPICH software to use Globus-family services. To improve performance, we use MPICH-based vendor implementations of MPI in local clusters.

CLUSTERIX has a hierachical architecture, with respect to both the memory access and communication. Inside SMP nodes, data exchanges between processors are performed through shared memory. SMP nodes are grouped into local clusters, and communications inside them are implemented using such network protocols as Gigabit Ethernet or InfiniBand. They are characterized by high bandwidths and small latencies (especially InfiniBand). Finally, local clusters connected by WANs are building blocks for the whole meta-cluster.

Taking into account the hierarchical architecture of the CLUSTERIX infrastructure, it is not a trivial task to adapt the existing applications for effective use in the meta-cluster. It requires parallelization on several levels corresponding to the meta-cluster architecture, taking into account the high level of heterogeneity in network performance between various subsystems of the meta-cluster.

3 Testing Applications

3.1 FEM Modelling of Castings Solidification

NuscaS is an object-oriented package for the FEM modeling, designed [7] at Czestochowa University of Technology to investigate thermomechanical phenomena. Its functionality includes also implementation on clusters. *NuscaS* is one of pilot applications adapted for the execution in the CLUSTERIX environment.

Fig. 1 present the performance results for the FEM modeling of castings solidification on two distant local clusters, located evenly in Czestochowa and Poznan. These results are rather promising. For example, the speedup for the mesh with 750313 nodes is $S_p = 15.05$, for $p = 18$ processors (9 from Poznan, and 9 from Czestochowa).

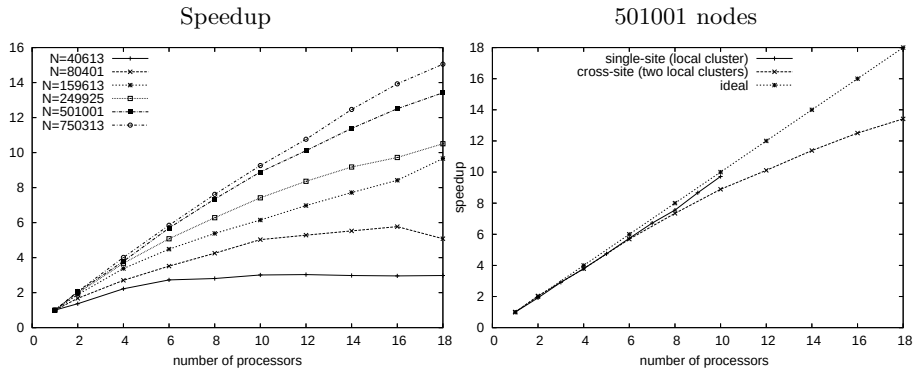

Fig. 1. Testing problem on two distant local clusters: (a) speedup for different mesh sizes versus number of processors; (b) speedup comparison of single-site and cross-site performance for mesh with 501001 nodes

3.2 Prediction of Protein Structures

Proteins are macromolecules which are absolutely necessary for functioning of all known living organisms. Each protein has a unique three-dimensional structure which determines in function. The theoretical prediction of the three-dimensional structure of proteins solely from its sequence is a grand challenge of computational structural biology and theoretical chemistry.

Successful application of protein structure prediction methods based on energetic criteria depends on both an adequate approximation of the energy function,

and an effective computational approach to global optimization of the conformational energy [6]. Prediction of the three-dimensional structure of protein using the conformational space annealing (CSA) as a global optimization method and the UNRES force field requires large computational resources [2]. The CSA algorithm is applicable to large-scale parallel computing, and in this work it has been extended to grid architecture of CLUSTERIX using the MPICH-G2 tool. Benchmark simulations show that CSA/UNRES can efficiently use distributed computing resources (see Table 1).

TASK		PB+PCz		TASK+PB+PCz	
p	time[s]	p	time[s]	p	time[s]
2	5394	1+1	5483		
4	1752	2+2	1837	2+1+1	2083
8	1722	4+4	777	4+2+2	1013
12	476	6+6	500	6+3+3	616
16	351			8+4+4	456
32	174			20+6+6	199

TASK+PB+PCz+WCSS		
p	time[s]	time[s]
12+0+0+0	476	
0+6+6+0	500	495
6+6+0+0	562	496
6+0+6+0	505	491
6+0+0+6		503
0+6+0+6		496
0+0+6+6		497
4+4+4+0	690	500
0+4+4+4		505
4+0+4+4		512
6+3+3+0	616	
3+6+3+0	643	
3+3+6+0	571	

Table 1. Perfmance results for single-site and cross-site execution of prediction of protein structures

References

1. CLUSTERIX Project Homepage, http://www.clusterix.pl
2. Czaplewski, C., et al.: Improved conformational space annealing method to treat β-structure with the UNRES force-field and to enhance scalability of parallel implementation. Polymer **45** (2004) 677-686
3. Globus Toolkit Homepage, http://www.globus.org/toolkit
4. Dong, S., Karniadakis, G.E., Karonis, N.T.: Cross-site Computations on the TeraGrid. Computing in Science & Engineering **7**, **5** (2005) 14–23
5. Karonis, N., Toonen, B., Foster, I.: MPICH-G2: A Grid-Enabled Implementation of the Message Passing Interface. Journal of Parallel and Distributed Computing **63**, **5** (2003) 551–563
6. Scheraga, H.A., Liwo, A., et al.: The protein folding problem: Global optimization of force fields. Frontiers in Bioscience **9** (2004) 3296-3323
7. Wyrzykowski, R., Olas, T., and Sczygiol, N.: Object-Oriented Approach to Finite Element Modeling on Clusters. Lect. Notes in Comp. Sci. **1947** (2001) 250-257
8. Wyrzykowski, R., Meyer, N., Stroinski, M.: Concept and Implementation of CLUSTERIX: National Cluster of Linux Systems. Proc. LCI International Conference on Linux Clusters: The HPC Revolution 2005, Chapel-Hill, NC, April 2005