

CADMOS

User policy for Blue Gene/P

Summary

This document aims to provide the information necessary for CADMOS to implement the user policy for the Blue Gene/P system (in short BG/P) in order to exploit its extraordinary computational power to the best advantage and allow the scientific community of the CADMOS partner institutions (UNIGE, UNIL and EPFL) to gain maximum benefit from it through the submission of suitable projects.

1 Characteristics of Blue Gene/P

The BG/P installed at the DIT/EPFL comprises 4 racks, each with 1024 nodes, in other words a total of 4096 nodes. A rack consists of two midplanes of 512 nodes. The midplane is the smallest part of the BG/P whose nodes may be interconnected in a three-dimensional torus (3D torus), i.e. the interconnection topology that makes Blue Gene so attractive compared to other parallel architectures. The nominal calculation performance is 56 Tflops.

A midplane may itself be subdivided into blocks of 128 nodes, and even of 32 nodes on one of the four racks. However, the nodes of such blocks cannot be interconnected in a 3D torus. These “small” blocks therefore make poor use of the BG/P interconnection capacity.

In order to execute a parallel application on the BG/P, a user asks the resource allocation system for the number of nodes that he wishes to implement. If available, the system allocates him the smallest free physical partition that can contain the number of nodes requested for his exclusive use and however long it takes for him to perform his task.

The BG/P system allocates predefined partitions of the machine in accordance with user requests. The partitions correspond to separate physical blocks of between 32 and 4096 nodes. Each partition can only be allocated to one single user at the same time.

This resource-sharing system therefore limits the number of tasks that can be performed simultaneously on the BG/P, especially if these tasks require a sufficient number of nodes to be described as massively parallel and benefit from a 3D torus interconnection topology. For example, a maximum of four tasks comprising 1024 nodes can be performed simultaneously on the machine.

1.1 Type of applications that can benefit from BG/P computational power

The BG/P is a massively parallel supercomputer, offering an extremely favourable correlation between processor computational speed and inter-node communication bandwidth. The machine is in fact said to be “balanced”, as its processor speed tends to be slow and its network bandwidth relatively large. The BG/P was thus essentially designed for parallel applications that can be efficiently executed on a very large number

of processors and require a large number of communications (frequent and/or voluminous).

1.2 Checkpoints and restart

Another aspect to be considered on the new CADMOS BG/P installation is the extraordinary capacity (approx. 1 Pbytes) and potential throughput (up to 10 Gbytes/sec) of the parallel file systems available to users.

On a system of this nature, it is essential that a large-scale simulation can be carried out in successive stages using the checkpoint and restart technique in order to

1. minimise the impact of a breakdown and
2. reduce the average task performance time.

Any massively parallel application running on the BG/P must therefore be capable of saving its status on disk (in a relatively short lapse of time in relation to the total execution time) and restarting from the most recent backup. The checkpoint/restart procedure must be provided to BG/P administrators.

2 Rules for the use of Blue Gene/P

The BG/P computer is installed at EPFL; the DIT/EPFL provides the services described in the document "*Service HPC sur Blue Gene*", which can be found on the CADMOS website.

The BG/P system is governed by the two modes of use :

- test mode and
- production mode,

the latter comprising several classes (see Table 1) :

- short,
- mid,
- large,
- and full.

The BG/P LoadLeveler task manager manages the priorities for all classes except *full*. The latter is managed directly by the BG/P system administrator.

Mode	Class name	No. nodes	Max Wall Time	Priority	Remarks
<i>test</i>	test	32,64,128	30 min.	10	
<i>production</i>	short	128,256	4 hours	10	
<i>production</i>	mid	512	12 hours	20	
<i>production</i>	large	1024, 2048	24 hours	20	
<i>production</i>	full	up to 4096	12 hours	Reservation	Specific request

Tab. 1: Modes, classes and priorities on BG/P.

2.1 Test mode

- For the *test* mode, the system offers 7 predefined small partitions (3x128 nodes and 4x32 nodes) on a dedicated midplane, in other words 1/8 of the total calculating resources, for 6 hours per day (or 12 or 24 if subsequently considered necessary), from Monday to Friday.
- To launch a task in test mode, BG/P administrators will set up an interactive batch system as quickly as possible. Meanwhile, users must choose a free test partition by consulting the site <http://bluegene.epfl.ch> and then execute the command `mpirun` interactively from a front-end machine.
- A test task is normally limited to 30 minutes in order to ensure equitable access to the partitions devoted to test mode.
- To obtain a test account on the BG/P, the following form must be sent to the local administrator of the partner institution concerned
“Inscription d'utilisateur(trice) aux serveurs de calcul du DIT-EPFL”
(<http://hpc-dit.epfl.ch/docs/inscription-utilisateurs-enligne.pdf>), with a brief description of the planned simulations (details of documents that must be provided can be found on the CADMOS website). This form will be validated by the local administrators, who will accredit users in the EPFL system and then ask the DIT/EPFL to open accounts.

2.2 Production mode

- BG/P is chiefly dedicated to production simulations, i.e. the carrying out of large-scale projects (cf. Section 3).
- In order to obtain a production account, a dossier must be submitted to the CADMOS Direction Committee (cf. Section 3, *Submission of a project*).
- The resources used are calculated in terms of “nodes * hours” of task execution. In case of saturation of the machine, the criterion of equitable distribution of resources between institutions must be taken into account.
- Within the limits of its resources, each partner institution can define relative priorities for its projects.
Furthermore, BG/P administrators can adjust the priorities on a monthly basis in order to balance the distribution of resources between the three partner institutions and, more generally, guarantee optimal exploitation of the system.
- Each partner institution can within the CADMOS Direction Committee impose the acceptance of projects it considers strategic, whilst respecting the balanced use of resources among partner institutions.
- With a view to promoting HPC in the social sciences, the UNDL Foundation also has access to BG/P.
- If a specific request is made to the CADMOS Direction Committee, exceptional allocations (typically a session devoted to debugging) of resources may be granted

- (exclusive access to all or part of the machine during a particular period). These resources are deducted in their entirety from the resources available for the project.
- Reservations may be altered by the CADMOS Direction Committee providing prior notice is given.
 - The LoadLeveler batch system manages the production midplanes.
 - LoadLeveler accepts parallel tasks requiring between 128 and 2048 nodes (see Table 1) .
 - The *full* class (up to 4096 nodes) is accessible only to projects that request it, in principle within the limits of the quotas of each partner institution. This request must be justified. To optimise the use of the BG/P, access to this class is scheduled in advance.

3 Submission of a project (production mode)

Only permanent scientific staff members of a partner institution may submit a project.

A project is submitted in 3 stages :

1. Tests on Blue Gene/P and local HPC resources (e.g. VitalIT, Callisto, possibly CSCS). This phase serves to “demonstrate” that the project’s parallel applications can be adapted to the large number of processors available on the BG/P.
2. 1 month maximum of trial on the *mid* queue (access must be requested from the DIT). Objective: show the scalability on several (minimum 2) platforms (e.g. Callisto and BG/P).
3. Submission of a scientific project request with :
 - brief description of scientific project,
 - description of paradigms used for parallelism,
 - scientific staff concerned and any existing projects already financed,
 - demonstration of necessity of using BG/P rather than another resource,
 - duration.

The decision will be made by the CADMOS Direction Committee on the basis of the scientific results obtained and anticipated and the status of requests for other projects. The Direction Committee will also decide on the duration, which will not exceed 12 months. A renewal may be requested via a simpler procedure.

The decision regarding whether or not to support a project is made by the CADMOS Direction Committee based on the following factors :

1. scientific quality of project and appropriateness of resources requested,
2. demonstration of usefulness of BG/P and its particularity in comparison with other architectures potentially available in Switzerland.

The allocation is defined in accordance with the project requirements and availability of the machine. Projects that are not accepted could be directed towards local HPC resources.

4 Use of Blue Gene/P

All activities using BG/P must have as their objective publication in scientific journals. Any dispensations must be granted by the Direction Committee.

The publications must accredit CADMOS resources as follows “*The financial support for CADMOS and the Blue Gene/P system is provided by the Canton of Geneva, Canton of Vaud, Hans Wilsdorf Foundation, Louis-Jeantet Foundation, University of Geneva, University of Lausanne, and Ecole Polytechnique Fédérale de Lausanne.*”

5 Transition period (until 31.12.2009)

Test accounts for CADMOS partners can already be requested, existing accounts (test and production) can be renewed until 31.12.2009 on request (email to Mr Christian Cléménçon, christian.clemencon@epfl.ch).

6 Conclusion

The CADMOS BG/P can undoubtedly be classed as a massively parallel computer. It nonetheless offers a limited number of computing partitions offering a very efficient interconnection network, thus restricting the number of tasks that can be performed simultaneously.

Given the nature and physical characteristics of the system, the use of large partitions (typically between 1024 and 2048 nodes) is given priority in production mode over applications that cannot utilise at least 512 nodes efficiently. The latter will be directed towards other parallel architectures, such as clusters, or a computing grid.

On the other hand, as far as the test and program tuning are concerned, experience with BlueGene/L has shown that it is advisable to make available to users a certain number of small partitions, if possible accessible interactively and without any waiting period.

7 Template for test mode

When sending a request in *test* mode, the following document must be completed and submitted together with proof of identity

<http://hpc-dit.epfl.ch/docs/inscription-utilisateurs-enligne.pdf>

This form will be submitted to the local administrator (at the addresses cadmos@epfl.ch, cadmos@unil.ch, or cadmos@unige.ch, depending on the institution concerned), who will accredit users in the EPFL system and then ask the DIT/EPFL to open accounts.

8 Template for production mode

The request for a project in *production* mode must be submitted in PDF format by email to the local administrator (cadmos@epfl.ch, cadmos@unil.ch, cadmos@unige.ch). The following information must be included :

1. Description of project (see Section 8.1),
2. CVs of applicants and any co-applicants,
3. A list of publications strictly relating to the project content. In the case of a renewal, include the list of any articles resulting from the current project.

Whenever possible, the CADMOS Direction Committee undertakes to reply within one month.

8.1 Description of project

The project must be submitted in the following form :

1. First page showing details of the project, including :
 - title of project,
 - name and institution of applicants (and co-applicants),
 - resources requested (see point 3 below),
 - estimation of typical number of processors.
2. Summary of project (1/2 page). For accepted projects, this summary will appear on the CADMOS website.
3. Description of project (max. 5 pages), including
 - background,
 - objectives, justification and description of research,
 - duration,
 - resources requested (estimated time, disk space necessary) and their justification,
 - scalability results on BlueGene/P and at least one other HPC system (after tests have been carried out),
 - anticipated results and scientific impact.