Usages et utilisateurs de Grid'5000: stratégie pour l'accès aux ressources

Lucas Nussbaum

▶ **To cite this version:**

Lucas Nussbaum. Usages et utilisateurs de Grid'5000: stratégie pour l'accès aux ressources. 2016. hal-01294910

HAL Id: hal-01294910
https://hal.inria.fr/hal-01294910

Preprint submitted on 30 Mar 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Usages et utilisateurs de Grid'5000
Stratégie pour l'accès aux ressources

Lucas Nussbaum, pour le
Comité des responsables de sites Grid'5000

Février 2016 - version de travail

Introduction

Le projet Grid'5000 a été initié avec l'objectif de servir les besoins en expérimentation (en particulier dans le domaine du HPC, initialement) de toute la communauté académique française. Toutefois, cela pose deux principaux problèmes :

— La non-différenciation entre les utilisateurs affiliés à des financeurs directs et les autres utilisateurs n'incite pas les autres utilisateurs à s'investir pour contribuer au financement de la plate-forme.

— Pour des raisons d'économie d'échelle et de rationalisation, il est logique pour les financeurs directs de Grid'5000 de chercher à intégrer Grid'5000 dans une stratégie de site globale à la problématique calcul et données ou du moins à acceuillir les besoins de calcul sur Grid'5000, notamment pour remplir les clusters ou les périodes horaires les moins utilisés. De plus, ces besoins sont souvent mal servis par les mésocentres (besoin de piles logicielles spécifiques, d'accès root, etc.)

  — Mais cela ne devrait pas se faire au détriment des besoins en expérimentation qui ont besoin de réserver des ressources précises pour leur expériences, et souvent de les utiliser de manière interactive (en interagissant avec les noeuds tout au long de la réservation).

Ce document propose une solution visant à remédier à cela. Il propose une classification de l'usage et des utilisateurs de Grid'5000 (Section 2), puis propose un ensemble de règles et de limitations pour réguler l'accès à l'instrument (Section 3).

Deux axes pour classifier l'usage et les utilisateurs

Les usages et utilisateurs de Grid'5000 peuvent être classés sur deux axes orthogonaux :

1. L'axe *usage* : l'objectif scientifique
   (a) Expériences en HPC, Cloud, infrastructures Big Data, réseau : l'objectif de l'utilisateur est d'utiliser Grid'5000 comme un *modèle* d'une plate-forme réelle, pour y instancier et y évaluer une solution (logiciel, framework). Les questions que se posent l'expérimenteur portent sur la performance, la faisabilité, souvent en comparant plusieurs approches ou solutions. Pour réaliser cette étude, un traitement ou un calcul est souvent fait, mais l'utilisateur ne s'intéresse pas au résultat du calcul ou traitement, mais plutôt à *comment* le calcul ou traitement s'est déroulé.
   Besoins :
   — Même s'il est possible d'automatiser le processus dans une certaine mesure, ce type d'usage nécessite souvent de nombreuses interactions entre l'expérimenteur et la plate-forme.
   — Les ressources matérielles utilisées pour l'expérience sont très importantes, puisque le résultat de l'expérience dépend directement du choix des ressources

   (b) Calcul de *production* : l'objectif de l'utilisateur est de réaliser un calcul, un traitement de données, ou une campagne de calculs ou de traitements. L'utilisateur s'intéresse au résultat du calcul, pas aux performances du processus de calcul lui-même (même si parfois, c'est un argument secondaire).
   Exemples :
   — Simulation P2P ou réseau avec SimGrid ou NS3 (car ici seul le résultat compte, pas la manière dont s'est déroulé le calcul)
   — Evaluation d'algorithmes d'évolution de robots par simulation (en intelligence artificielle / robotique)
   — Evaluation d'algorithmes de traitement d'images, de sons, de vidéos ou d'analyse de données
   Besoins :
   — Un grand volume de ressources
   — Ce type de traitements peut généralement être automatisé plus facilement que des expériences

— le choix précis des ressources n'est pas important même si : (1) c'est forcément plus intéressant si les ressources sont performantes $\rightsquigarrow$ calcul terminé plus vite ; (2) certains calculs nécessitent du matériel spécifique (réseau Infiniband, GPU)

2. L'axe de l'*affiliation institutionnelle* des utilisateurs, en fonction de leur niveau de contribution à l'infrastructure. En première approche, nous proposons une répartition en trois niveaux d'utilisateurs [1] :

Niveau "Gold" :

— Les équipes des centres Inria qui hébergent un site Grid'5000 (paiement des fluides) ou ont fléché des financements CPER vers Grid'5000 ou ont un scientifique local impliqué dans le comité des responsables de sites $\rightsquigarrow$ (au 1^{er} mars 2016) centres de Grenoble, Lille, Nancy, Rennes, Sophia (pas Bordeaux, Paris, Saclay).
— Les équipes des laboratoires (UMR) qui participent à l'hébergement d'un site, ou ont un scientifique local impliqué dans le comité des responsables de sites, ou fournissent un ingénieur à au moins 50% à l'équipe technique $\rightsquigarrow$ (au 1^{er} mars 2016) laboratoires CReSTIC (Reims), I3S (Sophia), IRISA (Rennes), IRIT (Toulouse), LIFL (Lille), LIG (Grenoble), LIP (Lyon), LORIA (Nancy), CSC (Luxembourg).
— Les utilisateurs payeurs (industriels)

Niveau "Silver" :

— Les équipes d'autres centres Inria (puisqu'Inria a contribué globalement via l'OIP et l'attribution de postes d'ingénieurs)
— Les équipes de laboratoires *voisins* des laboratoires *Gold* (par exemple, tutelle (université) commune ayant contribué au financement)

Niveau "Bronze" :

— Les autres utilisateurs académiques d'institutions françaises
— Les utilisateurs du programme Open Access

Cas particulier des collaborations de recherche
Un utilisateur qui serait impliqué dans un projet collaboratif (ANR, Europe) avec une équipe *Gold* ou *Silver* pourrait bénéficier de manière transitive de l'accès à l'instrument avec les privilèges de l'équipe *Gold* ou *Silver*.

Programme Open Access
Pour favoriser la visibilité internationale de la plate-forme en tant qu'instrumentation pour l'expérimentation en HPC/Cloud/Big Data/réseau, un programme « Open Access » a été mis en place pour permettre aux académiques étrangers d'essayer Grid'5000. Les comptes obtenus dans le cadre de ce programme ont une durée limitée (2 mois, éventuellement renouvelable). Nous proposons qu'ils soient classés au niveau *Bronze*.

Contrat pour l'utilisation de Grid'5000 par des industriels
Nous savons établir une convention d'accès à l'instrument contre rémunération (paiement au volume d'utilisation). Nous proposons que ces industriels, souhaitant expérimenter sur Grid'5000, soient classés au niveau *Gold*.

Proposition de règles pour l'accès à l'instrument

Nous proposons que chaque utilisateur soit catégorisé, d'une part par un type d'utilisation principal (expérimentation ou production), et d'autre part par son affiliation institutionnelle (Gold/Silver/Bronze). Cette classification serait implémentée dans l'outil de gestion de comptes Grid'5000 (UMS), probablement via la gestion des groupes d'utilisateurs.

En fonction de ces deux critères, chaque utilisateur aura accès (ou pas) à certaines *queues* (files d'attente), et des limites s'appliqueront aux jobs qu'il pourra y soumettre.

La soumission de tâches sur Grid'5000 pourra se faire via trois *queues* :

— **default** est la queue utilisée par défaut, destinée à recevoir l'utilisation « expérimentation » de Grid'5000. La plupart des ressources de Grid'5000 sont disponibles dans cette queue. Les jobs ne sont pas interruptibles. La charte fixe actuellement des limites sur le volume des jobs utilisés en journée (par jour et par cluster, pas plus que l'équivalent de deux heures sur tous les coeurs). Les réservations à l'avance sont possibles (mais avec un maximum de deux).
— **best-effort** permet de soumettre des jobs qui seront interrompus dès qu'un job dans une queue plus prioritaire aura besoin des ressources. L'avantage est qu'il n'y a pas de limitation sur le volume de ressources utilisées. Un outil (CiGri) permet de gérer automatiquement des grandes campagnes de jobs best-effort, en re-soumettant automatiquement les jobs interrompus. Toutes les ressources de Grid'5000 sont accessibles via cete queue.

1. Ce système de répartition ne prend pas en compte l'ampleur du financement, la date du financement (ni la manière dont un financement à un instant t serait *amorti* sur la durée), ou la proportionnalité entre le niveau de financement et le nombre d'utilisateurs bénéficiaires. Pour conserver des définitions simples (et éviter une approche extrèmement codifiée), le comité des responsables de sites serait chargé de définir annuellement la liste des institutions et laboratoires bénéficiaires des différents niveaux.

— **production** est une queue disponible uniquement sur des clusters du site de Nancy (en 2016), conçue pour être adaptée aux besoins de la charge de calcul de production. Seuls quelques clusters (qui ne sont pas accessibles via la queue **default**) sont accessibles via cette queue. Les réservations à l'avance sont interdites, il n'y a pas de limite de durée de job, mais les jobs sont ventilés sur les noeuds en fonction de la durée pour garantir qu'un job court démarrera après un délai d'attente court. Tous les utilisateurs de Grid'5000 peuvent y soumettre des tâches, mais on s'attend à ce que les utilisateurs hors LORIA / Inria Nancy – Grand Est utilisent leurs ressources de production locales pour les tâches qui ne nécessitent pas de fonctionnalités présentent uniquement sur Grid'5000.

Nous proposons de mettre en place les règles d'utilisation présentées dans le tableau 1. Seuls les utilisateurs *Gold* pourraient être catégorisés comme usage *production*, ce qui a pour effet de limiter l'accès à la queue *production* aux utilisateurs *Gold*.

	queue **default**			queue **best-effort**	queue **production**
	volume en journée (en cluster.heures)	nombre de réservations en avance par site	délai maximum avant la réservation		
expérimentation, Gold	2	2	sans limite	sans limite	oui mais moins prioritaire que (production, Gold)
expérimentation, Silver	2	1	48h	sans limite	non
expérimentation, Bronze	1	1	24h	sans limite	non
production, Gold	2	1	4h	sans limite	sans limite

TABLE 1 – Limites à l'utilisation pour chaque catégorie

Discussion

La limitation de volume en journée vise à assurer un partage des ressources pour permettre à un grand nombre d'utilisateurs d'accéder à la plate-forme en parallèle pour préparer leur expérience.

Les réservations à l'avance visent à permettre aux utilisateurs ayant besoin de ressources spécifiques de réserver ces ressources plusieurs jours (parfois plusieurs semaines) à l'avance pour leur expérience. Elles sont principalement utilisées pour les nuits et week-ends. La compétition est très rude pour réserver certains clusters à l'avance. L'objectif ici est de donner une chance aux utilisateurs (expérimentation, Gold) de réserver d'abord, puis aux utilisateurs (expérimentation, Silver), puis aux utilisateurs (expérimentation, Bronze), puis aux utilisateurs *production* (qui pourront ainsi récupérer les ressources qui n'ont pas été réservées précédemment).

L'usage de la queue *best-effort* resterait non régulé pour l'instant. Toutefois, dans le cas de compétition entre utilisateurs de la queue *best-effort*, une limite de volume pourrait être mise en place, au-dela de laquelle il serait obligatoire de passer par CiGri pour soumettre et permettre un arbitrage entre usagers de *desktop-computing* et campagnes de jobs best-effort.

La manière de rendre les usages (expérimentation, Gold) de la queue *production* moins prioritaires que les usages (production, Gold) reste à définir.

Limites de ce type de restrictions

Il faut noter que ce type de classification et de restriction ne prend absolument pas en compte l'intérêt et l'impact scientifique des travaux, et se limite à une approche pouvant être implémentée automatiquement (à la fois côté infrastructure pour la vérification de la conformité aux règles, et côté utilisateur pour maximiser le volume de ressources réservées avec un outil comme Funk).

Des pistes ont été évoquées dans ce sens (demander des rapports/slides aux gros utilisateurs sur leur utilisation de l'instrument, revue par une commission, puis attribution d'une priorité aux utilisateurs en fonction de leur impact scientifique). Pour l'instant, nous ne souhaitons pas avancer dans cette direction en raison du travail conséquent qui serait nécessaire pour arbitrer entre les utilisateurs de communautés très différentes.