Mutualisation des ressources de calcul parallèle à l'Université Louis Pasteur : un bilan

Romaric DAVID
Centre d'Études du Calcul Parallèle et de la Visualisation
UFR de Mathématique et d'Informatique, 7 rue René Descartes, 67084 Strasbourg
david@cecpv.u-strasbg.fr

Résumé

En menant une opération originale de mutualisation des moyens de calcul parallèle au sein de l'université Louis Pasteur de Strasbourg, le CECPV a proposé de nouveaux modes de redistribution et de partage de la puissance de calcul. Rassemblant plusieurs achats épars, cette opération a permis de doubler la puissance de calcul proposée à ses utilisateurs.

L'article présente la réalisation pratique de cette opération et ses différentes implications au niveau organisationnel. En particulier, nous détaillons les mécanismes innovants de partage des ressources mis en place et mettons en évidence les conséquences de la mutualisation sur la gestion du parc de machines. Nous examinons également les modalités du passage à l'échelle d'une telle opération, ainsi que son intégration dans une politique d'établissement.

Mots clefs

Calcul parallèle, Mutualisation, Partage de ressources, Retour d'expérience

1 Introduction

Le CECPV (Centre d'Études du Calcul Parallèle et de la Visualisation) est le centre de compétences de l'Université Louis Pasteur de Strasbourg (ULP) chargé de promouvoir le calcul parallèle et la visualisation dans les activités de recherche. À cette fin, le CECPV met gratuitement des ressources de calcul (plus d'une centaine de processeurs) à disposition de la communauté des chercheurs strasbourgeois. Les utilisateurs sont des équipes ou laboratoires disposant de codes parallèles de simulation numérique.

Il existe d'autres ressources de calcul parallèle à l'université : des groupes de recherche installent, gèrent et utilisent des clusters plus ou moins importants.

Partant de l'idée que cette dissémination des clusters peut nuire à une utilisation rationnelle des ressources humaines et des machines, le CECPV a mené une opération de regroupement de clusters en cours d'acquisition, afin d'agréger de la puissance de calcul et de la proposer à la communauté. Cette opération, dite de mutualisation des moyens de calcul parallèle, a permis de doubler la puissance de calcul disponible au CECPV tout en améliorant sa redistribution.

Dans cet article, nous présentons tout d'abord les différentes phases de l'opération. Nous nous intéressons ensuite aux procédures d'achat et aux nouvelles politiques d'exploitation induites par cette mutualisation. Par le biais d'indicateurs chiffrés, nous évaluons les résultats de l'opération. Enfin, avec plus d'un an de recul, nous pouvons tirer quelques enseignements de la mutualisation en examinant la question du passage à l'échelle et les critères selon lesquels un partenaire accepte ou non de participer à l'opération.

2 Aperçu de l'opération

2.1 Buts de l'opération

En mettant en oeuvre le projet de mutualisation, nous poursuivons deux objectifs :

- agréger de la puissance de calcul afin d'en rationaliser l'exploitation et en particulier de libérer du temps chercheur ;
- redistribuer doublement cette puissance de calcul : entre les équipes y ayant contribué financièrement d'une part, à l'ensemble des utilisateurs d'autre part.

2.2 Étapes de l'opération

L'opération de mutualisation vise à étendre les ressources de calcul offertes à la communauté par le CECPV. Fin 2005, le CECPV disposait d'un cluster unique, financé par le Contrat de Plan État-Région (CPER). Ce cluster (environ 400 Gflops de puissance crête -Linpack [1] -) était ainsi constitué :

- 30 bi-processeurs Itanium 2 (mono-coeur), équipés de 8GO de RAM ;
- espace disque partagé par NFS de 2×500GO ;
- réseau d'interconnexion Gigabit Ethernet (pour NFS) et Myrinet 2000 (2Gbits) pour les communications des codes de calcul (MPI).

Ce cluster, désigné par *cluster commun* dans la suite, constitue la principale ressource de calcul parallèle de l'université. Son taux d'utilisation moyen se situe entre 70%

et 80%, ce qui est comparable aux taux de charge des grands centres de calcul nationaux (IDRIS et CINES).

Le financement CPER se montait à 600 000€, incluant une contribution supplémentaire des collectivités locales (Communauté Urbaine de Strasbourg, Conseil Général du Bas-Rhin et Région Alsace). La mise en production a eu lieu en janvier 2004.

Un des principaux intérêts d'un cluster est sa facilité d'extension, noeud par noeud. C'est pourquoi, dès l'achat de la machine, nous avons proposé aux utilisateurs d'y ajouter des noeuds, en fonction de leurs possibilités budgétaires.

Jusqu'à début 2005, afin de maintenir la cohérence architecturale, nous envisagions uniquement des extensions sur une base de processeurs Itanium. Le coût unitaire du noeud (environ 10 000€), ainsi que l'intérêt croissant à ce moment-là autour du processeur Opteron peuvent expliquer le succès... nul de ces tentatives[1].

Afin de rendre réalisable ce projet de mutualisation, nous avons fait le choix de casser l'homogénéité du cluster en proposant d'ajouter des noeuds à base de processeurs Opteron (coût moindre -environ 4500€- pour une performance au moins équivalente[2], architecture bi-processeur intéressante). Après quelques mois de tractations avec plusieurs laboratoires, nous avons pu établir fin 2005 le premier plan d'acquisition. Le tableau 1 résume les quatre phases d'acquisition de processeurs Opteron, d'octobre 2005 à avril 2007 (cf. lexique à la fin de l'article)

Laboratoire / projet	***Financement***	***Noeuds***
FoDoMust	ACI	**3** (2 puis 1)
HouPic	ANR	**4**
IMFS	PPF Fonds propres	**3**
LBM	Fonds propres	**18** (4, 4, 10)
Observatoire	Fonds propres	**4** (2, 2)

Tableau 1: Opérations d'achat des clusters

L'opération a finalement permis de rassembler 32 noeuds bi-Opteron, munis de 4GO de RAM et d'une carte Myrinet double port, soit autant de noeuds que nous nommerons *mutualisés* que de noeuds *communs* (issus du financement CPER en 2003). La puissance de calcul est équivalente à celle du cluster Itanium selon les tests Linpack.

Une autre opération de mutualisation, en une seule phase, a conduit mi-2006 à l'achat de 17 noeuds Athlon 64 dual core, munis de 2GO de RAM et de deux interfaces Gigabit Ethernet pour les communications liées au calcul. Au titre du soutien à la recherche, le CECPV a proposé d'héberger et de gérer ce dernier cluster, destiné au développement d'une application de visualisation distribuée. Ce cluster est financé par l'ANR **Massim.** La puissance crête d'un Athlon correspond peu ou prou à celle d'un Opteron. Les trois clusters cohabitant au CECPV sont repris dans le tableau 2.

CPU	***Noeuds***	***Ram (GO)***	***Réseau comm.***
2×Itanium2	30	8	Myrinet
2×Opteron	32	4	2×Myrinet
Athlon dual	17	2	2×Giga-Ethernet

Tableau 2: Clusters présents au CECPV

Ces trois clusters sont construits autour de deux serveurs frontaux (un par classe d'architecture) et de deux serveurs NFS.

2.3 Procédures d'achat

Rappelons que le principe de l'opération est d'acheter des machines avec des budgets externes au CECPV. Pour ce faire, les procédures d'achat et de gestion des marchés sont prises en charge par le CECPV, qui en général avance les fonds puis refacture les ressources de calcul aux laboratoires.

Nous examinons l'évolution des différents éléments (machines, réseaux) intégrés à cette refacturation. En particulier, nous revenons sur les impacts du passage à l'échelle de l'opération en terme de coût final pour le laboratoire d'une part, pour le CECPV d'autre part.

Au début de l'opération (dans les 2 premières phases, concernant 12 machines Opteron), il était convenu ce qui suit :

- le CECPV fournit l'infrastructure réseau (ports de commutateur Ethernet et Myrinet) ;
- les laboratoires fournissent les machines équipées de leur carte réseau Myrinet + câbles + logiciel de gestion de file d'attente (LSF de Platform Computing dans notre cas).

L'infrastructure réseau étant déjà présente et suffisante, elle n'engendre pas de coût supplémentaire pour le CECPV. Le surcoût imputé aux laboratoires peut donc s'analyser comme un «coût d'intégration» aux backbones (épines dorsales) des réseaux existants. Ce surcoût correspond à 940€, soit 32 % du prix du matériel seul.

Ce mode de fonctionnement est parfait jusqu'au jour où les différents backbones s'avèrent sous-dimensionnés. Ceci

[1] À l'exception d'une mention dans un Programme Pluri-Formation (PPF) à la demande de l'ULP.

[2] Des tests avaient été réalisés, montrant des performances équivalentes ou 2 x supérieures suivant les applications.

s'est produit lors de l'ajout d'un bloc de 6 machines en milieu d'opération. Le budget du CECPV ne permettant pas de réaliser l'investissement lié à l'extension nécessaire des backbones (8000€), il a été décidé en accord avec les laboratoires contributeurs d'intégrer le prix du backbone au prix des machines. Ceci porte le surcoût, par machine, à environ 1500€. Ce montant important n'a – heureusement pour le projet – pas dissuadé les équipes d'acheter des machines.

Examinons à présent les modes de re-facturation mis en jeu. Le CECPV achète les machines, puis impute ce montant au laboratoire contributeur. Il faut distinguer deux cas :

- facturation vers un compte de l'université (prestation interne) ;
- facturation vers un compte externe (CNRS en particulier).

Toute recette sur un compte se traduit par le prélèvement de *frais de* gestion (12%). Fort heureusement, le transfert de fonds interne - depuis un autre compte de l'université - n'est pas soumis à ces prélèvements.

Un transfert de fonds interne peut également s'effectuer par le biais d'une décision budgétaire (DB) : une DB met immédiatement les fonds à disposition. L'inconvénient est la relative opacité de cette mesure (un simple commentaire dans l'application de gestion).

Les matériels acquis par ce processus, que la facturation soit interne ou externe, sont initialement propriété administrative de l'université. Le transfert de propriété entre deux composantes de l'université ne nécessite aucune formalité. Pour les achats financés par des composantes externes à l'ULP, seule la signature d'une convention de transfert de propriété permettrait de les sortir de l'inventaire de l'université.

Dans d'autres cas, nous avons pu faire payer directement un sous-ensemble d'une facture (clairement identifié) par l'entité finançant le matériel. Ce dernier fonctionnement est intéressant : il donne naturellement la propriété administrative du bien à l'acquéreur. Cette pratique est réalisable dans le cas ou le budget d'origine est étiqueté strictement *équipement.*

2.4 Utilisation des marchés de l'établissement

Un marché spécifique avait été mis en place pour l'achat des machines dans le cadre du CPER. Au contraire, les marchés informatiques de l'établissement ont été utilisés pour commander les clusters d'architecture x86-64. Le même fournisseur (moins-disant au niveau prix) a été retenu lors des différentes remises en concurrence. Les marchés de l'établissement nous ont donc permis, dans ce cas, de construire un cluster homogène sans mettre en place un marché à bons de commande spécifique. Il s'agit peut-être du seul avantage.

En effet, la dernière phase d'achat de noeuds *Opteron* a consisté en l'achat de matériel d'occasion (opportunité ponctuelle). Un marché spécifique a donc été mis en place, les marchés de l'établissement ne prévoyant que du matériel neuf. Les réponses à l'appel d'offres passé nous ont montré que le prix des composants Myrinet était bien moins élevé (-20%) que via les marchés de l'établissement, par le jeu des partenariats commerciaux entre constructeurs et revendeurs.

Lors de l'achat d'une partie de nos serveurs *Opteron*, nous avons donc payé trop cher nos cartes réseau, qui représentent presque 20% du prix des machines. Si nous avions pu prévoir dès le début le succès de l'opération, nous aurions gagné à mettre en place un marché spécifique à bons de commande avec un revendeur spécialisé machines de calcul.

Nous pouvons estimer la perte subie à 3600€, hors coût des commutateurs.

3 Politique d'exploitation

Nous rappelons le double but poursuivi dans le cadre du projet : agréger la puissance de calcul et la redistribuer à deux communautés distinctes.

Sur nos machines de calcul, les utilisateurs placent leurs travaux dans des files d'attente. Quand les ressources nécessaires (en particulier un nombre donné de processeurs) sont disponibles, le système de gestion des files d'attente place les travaux sur les noeuds du cluster désignés pour l'exécution.

Dans le cas d'une machine de calcul commune et homogène (un seul type de processeur), la politique d'exploitation qui régit le partage des ressources est relativement simple. Il s'agit de définir les classes de travaux : travaux fortement/faiblement parallèles, travaux longs ou courts. Chaque classe de travaux se voit rattachée à une file d'attente reprenant ces propriétés. Nous disposons de 3 classes :

- travaux parallèles moyennement longs (classe utilisée en production) ;
- travaux peu parallèles courts (utilisée durant les phases de développement de codes) ;
- travaux peu parallèles longs.

Dans chacune de ces files, auxquelles sont affectées *n* processeurs, les utilisateurs sont équi-prioritaires.

Ce modèle n'est pas adapté à une exploitation où certains utilisateurs des ressources en sont également les payeurs (nous les désignons utilisateurs contributeurs dans la suite). Nous garantissons à ces utilisateurs un accès immédiat et sans attente à leurs ressources.

Pour ce faire, nous avons défini des files préemptives sur d'autres. Les files préemptives (ou privées) sont ouvertes aux seuls utilisateurs contributeurs. Les files préemptables (ou publiques) sont ouvertes à tous. Ces deux types de files sont associées à l'ensemble des processeurs d'un cluster.

Lorsqu'un programme est soumis sur une file préemptive, le système de files d'attente réquisitionne au besoin les

processeurs d'un programme tournant au même moment sur une file préemptable. L'intégralité de l'application préemptée est suspendue jusqu'à terminaison du travail préemptif. Par exemple, dans la figure 3.1, les processeurs sont occupés de T1 à T2 par des applications « privées » (provenant d'utilisateurs contributeurs). Durant cette période, les codes « publics » ayant démarré à T0 sont endormis et reprennent à T2. Les files d'attente privées sont donc occupées de T1 à T2, les publiques dès T0 (un job suspendu ne sort pas de file d'attente, il y change simplement de statut).

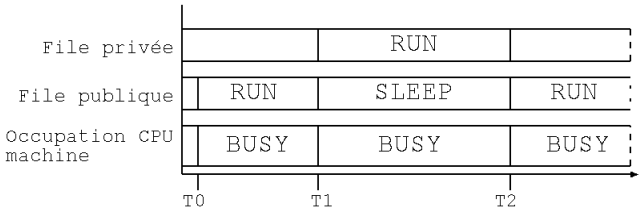

Figure 3.1: Occupation des processeurs

Ce mode de fonctionnement garantit aux utilisateurs contributeurs que, sans aucune latence, leurs applications seront prioritaires sur les applications publiques.

Comment réaliser à présent la redistribution des processeurs entre utilisateurs contributeurs ? Dans un premier temps (pendant un an), nous avons choisi d'instaurer une équi-priorité entre ces utilisateurs. Cela fonctionne bien dès lors que le nombre de machines financé par chacun des groupes est à peu près identique.

Dès lors qu'une grosse dissymétrie s'est présentée (un groupe disposant de la moitié des machines), nous avons mis en place une priorité proportionnelle au nombre de machines. Cette priorité garantit qu'en cas de très forte demande, une fraction du temps de calcul correspondante à la proportion du nombre de machines achetées sera restituée. On peut alors se demander quel est l'intérêt de la mutualisation pour les équipes contributrices. Un des éléments de réponse est que cela permet à une équipe ayant acheté *n* processeurs de lancer ses applications sur un nombre processeurs supérieur.

Un dernier mode de découpage serait d'attribuer de manière fixe *n* processeurs à une équipe ayant contribué pour la part correspondante. Ce fonctionnement présente le grand désavantage de laisser ces noeuds inoccupés en cas de baisse d'activité dans l'équipe.

4 Tableaux de bord

4.1 De nouvelles métriques

Le bouleversement de la politique d'exploitation nous a conduit à un constat d'insuffisance des indicateurs et tableaux de bord mis en place avant mutualisation. En effet, seul le nombre d'heures de calcul consommées triées par groupe d'utilisateurs nous intéressait. De même, le taux d'utilisation des ressources était calculé comme le rapport heures produites/heures productibles. En l'absence d'indicateurs sur le taux de disponibilité de nos machines, le nombre d'heures productibles correspond au nombre de processeurs × nombres d'heures écoulées dans la période considérée[3], pour le cluster donné.

Nous devons à présent :

- quantifier le nombre d'heures de calcul redistribuées à la communauté ;
- déterminer le nombre d'heures restituées aux contributeurs.

Pour un cluster donné, nous appelons *taux de mutualisation* le ratio heures redistribuées / heures productibles. De même, nous désignons par taux de restitution le rapport heures consommées par une équipe / heures productibles par les machines de cette équipe.

Nous étudions l'impact de ces indicateurs sur la politique d'attribution des ressources. Une des principales contraintes est de garantir à un groupe ayant contribué à *n* machines un taux d'utilisation de 100% sur ces *n* machines, si le besoin s'en fait ressentir.

Enfin, ces indicateurs nous permettent de caractériser les profils des opérations de mutualisation conduisant à une redistribution importante de puissance de calcul.

Nous avons ré-écrit l'application d'analyse des fichiers journaux du système de files d'attente afin d'intégrer le statut des files (privées, publique) et d'autres éléments (par exemple le nombre de processeurs) permettant de faciliter l'extraction des données.

4.2 Résultats

Nous présentons deux types de résultats :

- taux de mutualisation par cluster ;
- taux de restitution avant et après modification des priorités intra-file.

Rappelons que le taux de mutualisation est défini par le ratio heures produites sur files mutualisées / heures productibles.

Cluster	***Heures***	***Taux***
Opteron	2940	0,5
Athlon	147023	45,0

Tableau 3: Taux de mutualisation

Le tableau 3 nous montre deux tendances radicalement différentes. Le faible taux de mutualisation sur le cluster Opteron s'explique par l'utilisation qui en est faite : il s'agit d'une machine de production dont les files d'attente privées sont pleines en permanence.

Le cluster Athlon est utilisé pour une activité de développement de codes, donc sporadiquement. De plus, ces ma-

[3]L'habitude est de considérer que tous les mois de l'année font 30 jours.

chines ont bénéficié de plus de publicité. Le taux de mutualisation est donc logiquement plus élevé.

Si la redistribution vers les *non-contributeurs* n'est pas toujours un succès, qu'en est-il de la redistribution *intra-contributeurs* ?

Pour le savoir, nous regardons, groupe par groupe, le *taux de restitution*, c'est à dire le ratio nombre d'heures consommées/nombre d'heures productibles sur les seules machines financées par ce groupe. Nous reportons également le maximum du taux de restitution sur un mois (tableau 4 : mode équi-prioritaire, tableau 5 : mode proportionnel, cf. point 3).

Groupe	***Heures***	***Rest. moyenne***	***Rest. max***
LBM	44948	74	116
IMFS	24719	52	75
Observatoire	6394	20	80

Tableau 4: Restitiution en mode équi-prioritaire

Un taux maximum supérieur à 100% signifie que sur un mois, le groupe a consommé plus de ressources que celles auxquelles il avait contribué - c'est aussi un des bénéfices de la mutualisation -.

Groupe	***Heures***	***Rest. moyenne***	***Rest max***
LBM	16174	71	86
IMFS	12063	139	159
Observatoire	5330	93	121

Tableau 5: Restitution avec priorités proportionnelles au nombre de machines

À l'heure où l'article est écrit, nous avons peu de recul sur ces chiffres. Une première explication à l'augmentation du nombre de taux de restitution supérieurs à 100% serait la conjonction de l'augmentation des besoins des groupes *IMFS* et *Observatoire* associée à la difficulté ponctuelle du groupe LBM à faire passer une application suite à une mise à jour système...

De plus, au sein d'un groupe, les différents travaux sont placés dans une file (FIFO, premier arrivé, premier servi). Selon le nombre de processeurs demandés, cet ordonnancement n'est pas forcément le meilleur. Il peut laisser des *trous* dans le nombre de processeurs utilisés, remplis par d'autres travaux en provenance d'autres groupes.

4.3 Enseignements et bénéfices

Nous pouvons dresser le bilan suivant. Premièrement, **le type d'activité** effectuée sur les ressources de calcul influe sur la quantité de redistribution effective. Ainsi, une machine dédiée au développement se prête bien à la redistribution extra-contributeurs, au contraire des machines utilisées en production d'heures de calcul.

L'intérêt de l'opération réside dans ce cas dans le meilleur partage intra-contributeurs, sans bénéfice pour le reste de la communauté. Si l'on souhaitait que les non contribteurs puissent bénéficier de ces machines ; il faudrait par exemple leur assurer un droit d'utilisation minimum dans les files d'attente correspondantes, au détriment des contributeurs. Reste à savoir comment les laboratoires ayant financé les ressources accepteraient cet *impôt CPU*...

Si l'établissement dans son ensemble a pu bénéficier de puissance de calcul *gratuite* et de mutualisation du temps d'ingénieur, les laboratoires contributeurs ont également réalisé des économies :

- pas d'investissement en climatisation (pris en charge par le CECPV) ;
- pas d'investissement en alimentation électrique (onduleurs existants assez puissants) ;
- pas de transformation de temps chercheur en temps ingénieur. Les dernières mesures d'évaluation des chercheurs sont sans doute une incitation à valoriser encore d'avantage ce temps chercheur.

5 Passage à l'échelle

5.1 Aspect informatique

L'opération de mutualisation a dépassé le stade de simple maquette et nous a demandé de revoir totalement le dimensionnement de nos équipements.

Les commutateurs dédiés aux communications des codes de calcul sont construits autour de fonds de paniers dans lesquels des modules de 8 ports peuvent être ajoutés. Les premières phases de mutualisation (12 machines) ont pu s'appuyer sur un commutateur existant.

La phase suivante (6 machines, fin 2006) nous a conduit à étendre les capacités du réseau d'interconnexion. Face aux prix de celui-ci, nous avons envisagé un changement de technologie (Myrinet vers Infiniband). Le coût au Gbit/s était à ce moment-là en faveur d'Infiniband. Il aurait néanmoins fallu ré-équiper en Infiniband les 12 machines déjà acquises. Au frais de quelle entité ? De plus, ceci n'était rentable que pour des commutateurs Infiniband de taille trop limitée par rapport aux prévisions d'extension du cluster mutualisé. Nos prévisions ciblaient 32 machines. Nous avons finalement fait le choix de conserver Myrinet, qui permettait de plus une progression linéaire des investissements en réseau. Enfin, cela nous permettait de réutiliser au mieux un commutateur Myrinet existant et de refacturer le moins possible aux contributeurs.

5.2 Aspect gestion

La gestion d'un cluster est d'autant plus simple que le matériel est homogène. Le cluster d'Opterons, constitué en 4 phases, comporte 2 types de machines différents, ne se distinguant vu du système d'exploitation que par des détails comme la connectivité disque. Concernant les machines neuves, le suivi des références chez notre fournisseur nous a permis d'acheter le même modèle à 1 an d'intervalle.

Lorsque la dernière phase de l'opération (à base de machines d'occasion) s'est présentée, nous avons mesuré le risque d'augmentation d'hétérogénéité que nous prenions,

qui s'est avéré très raisonnable compte tenu du prix des machines.

Nous avons défini dans le point 4 différents indicateurs utilisant le nombre de machines des différents clusters. Ce nombre étant variable dans le temps, nous devons décomposer le calcul des différentes moyennes les mois durant lesquels des machines ont été ajoutées. Une application automatisée d'extraction de ces taux devrait donc utiliser un nombre moyen de processeurs par mois, en fonction de la date d'entrée en service des différents processeurs. Notre application ne permet pas encore d'automatiser cette tâche. Nous travaillons donc sur des mois pleins, sans tenir compte des mois durant lesquels le nombre de processeurs à changé.

5.3 Aspect logistique

La mutualisation des ressources de calcul se traduit par leur hébergement dans la même salle machine. En effet, l'existence du réseau local d'interconnexion haut débit, qui implique quantité de câblage, nous a obligé à regrouper les machines. Les murs de la salle ne sont évidemment pas extensibles à l'infini. En fin d'opération, l'encombrement au sol, tout comme la quantité de chaleur dégagée, limitent clairement l'extension dans sa conception mono-salle.

Si les locaux venaient à manquer, nous pourrions envisager la mise en place d'un ensemble de salles machines réparties sur plusieurs campus, administrables à distance (administrables avec KaDeploy [2] et IPMI [3] comme ce qu'on trouve par exemple sur les machines du projet Grid 5000 [4]). Ceci serait particulièrement intéressant pour des clusters construits à base de machines *grand public*, sans réseau d'interconnexion spécifique. Il faudrait alors mettre en place des systèmes de files d'attente multi-sites.

6 Étude des partenariats

6.1 Quand ça marche...

Analysons les raisons qui ont poussé les équipes à contribuer au projet, ainsi que celles qui ont poussé d'autres équipes à ne pas contribuer...

Le LBM est une équipe de recherche dépourvue d'informaticien attitré. Pour cette équipe, externaliser les machines est intéressant en terme d'économie de temps chercheur. L'équipe a parfaitement compris que son besoin était « un besoin d'heures CPU, pas de machines »

L'IMFS a été incitée fortement par l'Université à co-signer un PPF concernant le financement de machines. En effet, l'IMFS et le CECPV, de manière non concertée, avaient déposé 2 PPF dans lesquels figuraient des demandes de moyens de calcul. L'université a donc demandé le regroupement de ces deux dossiers.

Des collaborations scientifiques ayant été engagées entre le CECPV et l'Observatoire, il paraissait naturel à cette équipe de contribuer au projet. Leur motivation portait sur la redistribution intra-contributeurs et sur les caractéristiques techniques des machines. Bien que cette équipe dispose d'ingénieurs attitrés, l'aspect scientifique des collaborations antérieures a permis la mutualisation.

Le projet FoDoMust a été raccroché *in extremis* à l'opération de mutualisation. Démarchée tardivement. L'équipe a accepté de contribuer au projet pour les raisons suivantes :

- intérêt de principe pour le projet ;
- manque de moyens logistiques pour héberger les machines ;
- certitude de pouvoir extraire *ses* machines du cluster mutualisé à tout instant, par exemple en cas de désaccord sur l'exploitation.

Ne disposant ni de locaux ni d'ingénieur dédié, le projet Massim a apporté le cluster *Athlon* au projet.

Enfin, le projet Houpic (4 noeuds Opteron) s'est associé à la mutualisation pour bénéficier de la redistribution intra-contributeurs (accès à un nombre de processeurs supérieur à ceux achetés).

6.2 Et quand ça ne marche pas

L'article présente le projet de mutualisation initié en 2005. Nous avions déjà entamé des démarches antérieures, qui s'étaient révélées infructueuses.

Parmi celles-ci, nous avons contacté une équipe disposant d'un ingénieur attitré, et d'une solide culture de l'autonomie. Nous leur avions proposé de s'associer à nous pour l'achat de modules CPU s'intégrant dans une machine parallèle monolithique, puis pour l'achat de noeuds de cluster quelques années plus tard. Ces deux tentatives ont été couronnées d'échec.

La culture d'autonomie est très forte à l'université. Nous avons suggéré à une équipe de recherche *autonomiste* ne disposant pas d'informaticien attitré et souhaitant acheter des machines identiques aux nôtres de se joindre au projet. Notre suggestion est restée lettre morte.

Enfin, comme dernière illustration, un projet de machine de calcul mené par un laboratoire muni d'une équipe d'informaticiens conduira à l'installation de ressources de calcul découplées de celles du CECPV.

6.3 Politique d'établissement et mutualisation

La mutualisation des ressources de calcul parallèle nous paraît être un outil de gestion rationnel :

- du temps ingénieur ;
- de la puissance de calcul ;
- des coûts logistiques informatiques de l'université.

Pour continuer le projet, nous aurions besoin du soutien renforcé de l'université à plusieurs niveaux. Tout d'abord, pour simplifier la gestion comptable des opérations de refacturation, nous souhaiterions que les recettes correspondant à la mutualisation ne soient pas imposées. Dans le but d'inciter les laboratoires à contribuer, nous proposons que l'université prenne en charge le raccordement à l'infrastructure réseau. Enfin, et cela a été fait en partie dans le cadre de ce projet, nous souhaiterions que l'université continue de prendre en charge les coûts logistiques (climatisation, électricité).

Un projet comme la mutualisation a toutes les chance d'échouer dans des cas où les personnels sur place pourraient

se sentir dépossédés d'une partie de leur travail. L'envie (ou le besoin ?) de *maîtrise* des moyens informatiques peut aussi empêcher l'acceptation de leur externalisation.

7 Conclusion

Notre opération de mutualisation nous a montré que la redistribution des ressources de calcul était effective, tant au niveau intra-contributeurs que vers l'ensemble de la communauté. Pour le CECPV, cela permet de valoriser le projet au delà d'une simple opération d'hébergement de machines. Un des autres intérêts est également de favoriser les collaborations scientifiques qui sont primordiales dans nos missions.

Le passage à l'échelle ou la pérennisation d'une telle opération ne peut selon nous s'envisager sans incitation de l'université envers les acquéreurs. Des solutions innovantes de gestion logistiques devront alors être imaginées afin d'éviter l'engorgement des salles machines.

Lexique

Groupes, laboratoires ou projets ayant participé à l'opération :

- FoDoMust (Fouile de Données Multi-stratégie) : ACI dirigée par P. Gançarski. ;
- HouPic : (High Order Finite Element Particle-In-Cell Solvers on Unstructured Grids) : Projet ANR porté par E. Sonnendrücker
- IMFS : Institut de Mécanique des Fluides et des Solides. Nous travaillons avec l'équipe de J. Dusek
- LBM : Laboratoire de Biophysicochimie Moléculaire, dirigé par R. Stote
- MASSIM (MASses de données complexes de grande taille issues de la SIMulation numérique) : projet ANR porté par J-M. Dischler
- Observatoire astronomique de Strasbourg. Nous travaillons avec l'équipe dirigée par O. Bienaymé (puis R. Ibata).

Bibliographie

[1] J. Dongarra Performance of Various Computer Systems Using Standard Linear Equation Software, version **mise à jour en 2006** de P*roceedings of the Third Conference on Multiprocessors and Array Processors*, San Diego, California, Janvier 1987

[2] Yiannis Georgiou et al., A tool for environment deployment in clusters and light grids, Actes du Second Workshop on System Management Tools for Large-Scale Parallel System, Rhodes Island, Greece, Avril 2006

[3] IPMI : Intelligent Platform Management Interface, www.intel.com/design/servers/ipmi/

[4] Emmanuel Jeannot, Philippe d'Anfray, GRID'5000 une plate-forme d'expérimentation pour les systèmes distribués à large échelle, Actes de la conférence Jres 2007