PlaFRIM
Plateforme Fédérative pour la Recherche en Informatique et Mathématiques

Réunion du comité des utilisateurs

2 Mai 2016

SOMMAIRE

1. Arrêt de PlaFRIM1
2. Animation scientifique
3. Politique de mise à jour des systèmes, logiciels, bibliothèques, ...
4. Discussion sur le système d'ordonnancement (queues de soumission, limitations)
5. Présentation des machines en prêt

1

Arrêt de PlaFRIM1

Cas de manumanu

- STORM : utilisation de la machine, mais accès au lustre et aux modules partagés non nécessaires
- TaDAAM :
  - Intéret de faire faire la mise à jour par SGI par rapport à la faire nous-même ? Le support des UV est intégré dans Linux depuis longtemps. Il risque juste de nous manquer des outils propriétaires SGI. Mais si ca se trouve on a les licenses pour les réinstaller ensuite ? Et même, de quels outils SGI se sert-on ? Le MPI de SGI ? Des outils d'analyse de perfs ?
  - Tester avec un live CD de CentOS sans tout casser ? Utiliser une partition libre pour faire un dualboot temporairement ?
- LFANT :
  - Proposition d'utiliser la FRM pour le coût de mise à jour de la machine
  - Passage sous GUIX

Autres machines

- Les fourmis et les mirabelles seront migrées au cas par cas. Du fait de problèmes accumulés de versions de BIOS et IPMI incompatibles ou non avec la nouvelle version du noyau Linux, la migration doit se faire machine par machine, et prend du temps.
- Les mirages sont déjà sous PlaFRIM2.

Raisons de l’arrêt

- Machines vieilles de plus de 4 ans (peut être même 5 pour l’ensemble des machines)
- Sécurité obsolète de PlaFRIM 1 par rapport à PlaFRIM 2
- Equipe PlaFRIM en sous effective. Une première tentative de migration avait été faite et avait nécessité beaucoup de man power.
- Actuellement les modules ne sont plus mis à jour.

Remarques de l’équipe MEMPHIS

- Besoin d’une queue très longue comme celle des mirabelles.
- Outils Monika et Ganglia disponibles sur PlaFRIM 1 et pas réellement remplacés par des équivalents sur PlaFRIM 2
- Problème suite à la migration des mirabelles et fourmis sur PlaFRIM2
- Mirabelles dédiées à la formation ?
- Problème des « nodes failed » et problèmes d'écriture sur PlaFRIM2 qui arrivent de façon fréquente conduisent les utilisateurs à se rabattre sur PlaFRIM1 qui est toujours très sollicitée
- Problème de documentation et d'information (changement des modules et des variables d'environnement) sur PlaFRIM2
- Problème de quotas par équipe, des utilisateurs ne peuvent pas calculer sur PlaFRIM2 : les utilisateurs commencent à calculer sur PlaFRIM2, devant l'urgence des résultats et l'instabilité de PlaFRIM2 calculent sur PlaFRIM1

2

Animation Scientifique

Animation Scientifique

- Journée scientifique prévue avec le mésocentre.

  Organisation en cours. => sensibiliser vos équipes à y participer.

- Suggestions ?

3

Politique de mise à jour des systèmes, logiciels, bibliothèques, ...

Mise à jour des systèmes, logiciels, bibliothèques, ...

- Maintenus par l' équipe technique
- Modules disponibles dans /cm/shared/modulefiles
- Politique de mise à jour : on garde au minimum 3 versions :
  - standard
  - stable
  - Expérimentale

  GCC : compiler/gcc/4.8.4 compiler/gcc/4.9.0 compiler/gcc/4.9.2 compiler/gcc/5.1.0 compiler/gcc/5.3.0

  MPI Intel : mpi/intel-mpi/64/4.1.3/049 mpi/intel-mpi/64/5.0.3/048 mpi/intel-mpi/64/5.1.1/109 mpi/intel-mpi/64/5.1.3/181

Mises à jour système

- Patch de sécurité ASAP
- Mise à jour de la distribution linux
  - quand, comment ?
  - faut-il garder une image
    - Logiciel (libc, …) :
    - Noyau
- Comment rejouer une expérience lancée il y a 1 an ?
- Procédure pour la mise à jour, pour revenir en arrière ?

4

Discussion sur le système d'ordonnancement (queues de soumission, limitations)

Noeuds Miriel[001-077]

Queue	Memory	CPU Time	Walltime	Node	Cores	Max User Running	Max User Queuable	Max Job Running	Max Job Queuable
defq	-	-	< 02:00:00	< 4	<4	15	30	-	30
court	-	-	< 04:00:00	< 42	< 1008	2	10	20	64
longq	-	-	< 72:00:00	< 42	< 1008	2	10	16	64
special	-	-	< 00:30:00	< 77	< 1848	10	20	40	50

Noeuds Mistral[01-18]

Queue	Memory	CPU Time	Walltime	Node	Cores	Max User Running	Max User Queuable	Max Job Running	Max Job Queuable
court_mistral	-	-	< 04:00:00	< 18	< 360	2	10	20	10
long_mistral	-	-	< 72:00:00	< 16	< 320	2	5	10	10

Noeuds sirocco[01-05]

Queue	Memory	CPU Time	Walltime	Node	Cores	Max User Running	Max User Queuable	Max Job Running	Max Job Queuable
court_sirocco	-	-	< 04:00:00	< 5	< 120	2	4	10	5
long_sirocco	-	-	< 72:00:00	< 2	< 48	1	2	4	2

Noeuds souris

Queue	Memory	CPU Time	Walltime	Node	Cores	Max User Running	Max User Queuable	Max Job Running	Max Job Queuable
court_souris	-	-							
long_souris	-	-							

Queue **long_souris** réservée aux équipes LFANT et geostat

- Machine acquise pour les besoins spécifiques de ces équipes
- 3To de RAM pour jobs massifs de moyenne durée (**1 semaine**)
- Possibilité d'offrir ponctuellement ce même type d'utilisation pour d'autres équipes sur simple demande à **plafrim-support**

Queue special et multiPart

Queue « testpreempt » (expérimentale)

- Queue permettant de soumettre des jobs sur tous les nœuds libres
- ATTENTION: Les jobs sont soumis à préemption :
  - Si un des nœuds utilisé est requis par une autre queue classique, je job est détruit puis relancé automatiquement ultérieurement pour le temps restant
  - Importance de faire sois même du *checkpointing* (sauvegarde de l'état du calcul) régulièrement (toutes les heures ou ½ heures par exemple) pour pouvoir reprendre les calculs
- Limite actuelle en temps : 8 heures

HiePACS : Gestion des queues et cohabitation entre jobs longs et jobs courts

- Deux types d'utilisations de la plate-forme cohabitent.
- "Problèmes" lorsque ces deux modes d'utilisation (jobs très long et jobs très courts) visent les mêmes noeuds (typiquement les noeuds à beaucoup de coeurs de beaucoup de mémoire).
- (Re-)Discussion sur la stratégie mise en place pour la gestion des queues et des jobs
  - Ne faudrait-il pas par exemple comme c'est le cas dans certains centres de calcul limiter le nombre de jobs autorisés dans le système pour un utilisateur donné ?
  - Ne faudrait-il pas remettre à plat les métriques utilisées pour des mécanismes tels que le fair-share pour améliorer les choses ?

STORM : Jobs courts

- Dans les phases de développement, on a besoin d'un sous-ensemble de ressources, voire de l'ensemble, sur une période très courte, environ 10-15 mns.

MEMPHIS

- Mistral : machines ayant des carte XeonPhi obsolète =>disposer d'une queue 1 semaine max sur 8 ou 16 nœuds
- mirabelle : les conserver pour le moment dans PlaFRIM1 dans leur état actuel, queue de 1 mois
- miriel : avoir une queue - bloquée par défaut – permettant de soumettre des jobs de plus de 15 jours sur 8 à 10 noeuds sur demande motivée à plafrim-support

5

Présentation des machines en prêt

IBM POWER8

- bi socket
- 10 coeurs par socket
- supporte le multi-threading (smt1, smt2, smt4, smt8) donc 1,2,4 ou 8 threads par coeurs
- une bande passante vers la DRAM d'environ 300 Go/s
- 2 cartes graphiques K40

IBM Power8


Machine (255GB)
Group0 (128GB)
NUMANode P#0 (64GB)
Socket P#0
PCI 1014:034a
sda
sdb
sdc
sdd
sde
sdf
sdg
sdh
sr0
Core P#32
Core P#40
Core P#48
Core P#96
Core P#112
NUMANode P#1 (64GB)
Socket P#1
PCI 10de:1023
card0
PCI 14e4:168a
eth0
eth1
eth2
eth3
Core P#168
Core P#176
Core P#224
Core P#232
Core P#240
Group0 (128GB)
NUMANode P#16 (64GB)
Socket P#16
PCI 14e4:168a
eth4
eth5
eth6
eth7
Core P#2088
Core P#2096
Core P#2144
Core P#2152
Core P#2160
NUMANode P#17 (64GB)
Socket P#17
PCI 10de:1023
card1
Core P#2208
Core P#2216
Core P#2272
Core P#2280
Core P#2288
Host: power8.plafrim.cluster
Indexes: physical
Date: Mon 14 Mar 2016 02:59:38 PM CET

White Box KNL

Arrivée prévue en Avril

- un seul noeud ...
- 72 coeurs
- supporte l'hyperthreading
- unités vectorielles (2 VPU par coeur)
- support AVX512
- présence de MCDRAM (16Go) pouvant être utilisée comme cache ou comme mémoire propre (hiérarchie multiple et modulaire)
- une bande passante vers la DDR4 d'environ 400 Go/s
- pas de support OmniPath pour le moment ….

Merci !

Inria

Olivier Coulaud
Guilhem Savel
Hervé Mathieu

François Rué
Aurélien Dumez

LaBRI

Pascal Ung
Nathalie Furmento

IMB

Philippe Depouilly

Laurent Facq