Frigid'R : salle machine sans clim, extrême freecooling

1 Calculateur pour la grille de CIMENT (CiGri)

CIMENT [1] est le mésocentre de calcul intensif de l'Université Joseph Fourier. Il organise un réseau d'experts qui maintiennent plusieurs supercalculateurs utilisés par l'ensemble de la communauté scientifique Grenobloise. Chaque calculateur est en général lié à un ou plusieurs projets, mais CIMENT assure le fonctionnement d'une grille (CiGri) qui exploite les ressources des différents calculateurs quand elles sont libres, selon un principe de « meilleur effort » (c'est-à- dire lorsqu'aucune tâche locale prioritaire ne tourne dessus). La grille dans sa totalité est accessible à tous les utilisateurs de CIMENT, quel que soit leur laboratoire de provenance.

Jusqu'à maintenant, aucune machine n'était entièrement dédiée à la grille. Mais afin d'inciter les utilisateurs à mieux exploiter la grille chaque fois que cela est possible, CIMENT a décidé de se doter d'une machine qui sera accessible en priorité en mode grille. Le financement de cette machine est fourni par des contributions de différents laboratoires motivés par le projet. Les ressources ainsi acquises s'ajoutent aux ressources actuelles de la grille à la différence près que ces ressources n'ont pas la contrainte du mode « meilleur effort ». Le premier contributeur est l'Observatoire de Grenoble qui a acquit 28 nœuds de calcul bi-xeon avec des processeurs low-power. Cette première tranche représente 336 cœurs et 2To de RAM totale, avec un réseau d'interconnexion à faible latence Infiniband, pour une consommation électrique d'environ 5kW.

2 Refroidissement par « free cooling »

L'intégration d'une machine exploitée par CiGri et les problématiques actuelles de l'hébergement (coût du refroidissement en particulier, mais aussi éco-responsabilité) nous ont amenés à penser que cette machine pourrait fonctionner avec un refroidissement 100% par air ambiant, avec une circulation d'air appropriée.

Dans la région de Grenoble, la température extérieure est suffisamment basse pour permettre un refroidissement direct d'un supercalculateur durant 85% de l'année en moyenne en considérant une température ambiante de fonctionnement de 25°C (voir l'expérience du LPSC « Ecoclim » [2]). Ce pourcentage augmente avec des matériels de classe ASHRAE [3] supérieure ou égale à 2, ce qui est possible pour la partie « nœuds de calcul » d'une machine parallèle (température ambiante de fonctionnement pouvant aller jusqu'à 35°C).

Etant donné que les ressources disponibles dans une grille sont vues de manière globale, et que les taches sont exécutées en mode « batch », il est acceptable d'imaginer qu'une partie des ressources soient indisponibles pendant les périodes de l'année où il fait le plus chaud (par exemple, la machine peut fonctionner la nuit au mois de juin et être éteinte en journée).

Nous avons donc mis en place une solution légère destinée au refroidissement de ce nouveau calculateur selon ce mode, mais dimensionnée pour évoluer en fonction d'autres besoins qui pourraient s'y attacher. Le calculateur en particulier, est acheté en plusieurs tranches afin de monter en charge progressivement. D'autre part, à l'antenne du LIG [4] de Montbonnot (bâtiment INPG), certains serveurs expérimentaux bénéficient de ce refroidissement (par exemple : serveurs « bac-à-sable », forge en cours de développement, serveurs de secours pouvant être déplacé physiquement en cas de besoin,...). En conséquence, cette installation free-cooling a permit d'éviter d'investir dans une mise à jour du climatiseur qui était sous-dimensionné et pour lequel il aurait fallu prévoir un budget plus conséquent.

En outre, les équipes qui travaillent dans ces locaux (MESCAL et MOAIS) développent des outils de gestion de ressources qui sont utilisés entre autres dans CIMENT. CiGri fait partie de ces outils et ce logiciel fait souvent l'objet de recherches en ordonnancement. La prise en compte des contraintes du

free-cooling est quelque chose de très intéressant sur le plan expérimental et donne lieu à des développements innovants.

3 Une collaboration CIMENT/LIG exemplaire

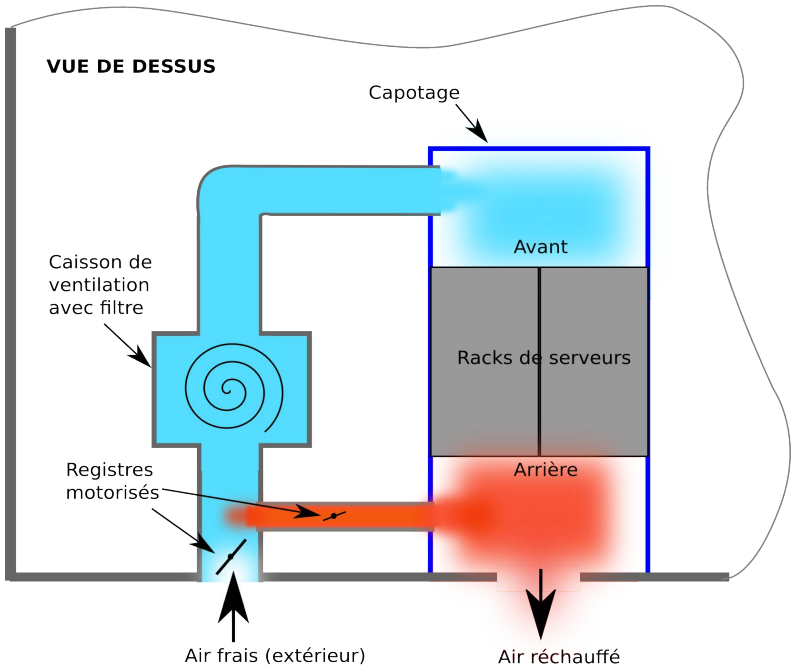
La collaboration entre CIMENT et les équipes MESCAL/MOAIS du LIG (Laboratoire d'Informatique de Grenoble) ne date pas d'aujourd'hui ! Dès sa création en 1998, CIMENT a apporté les cas d'usage et des ressources, et les équipes MESCAL/MOAIS ont fourni les outils qui sont issus de la recherche en informatique distribuée. Depuis plusieurs années, la gestion et l'optimisation de l'énergie fait partie des travaux menés en commun.

Ce projet, qui se situe dans une thématique d'actualité qui est l'éco-responsabilité en matière d'hébergement de matériel informatique énergivore, est dans la continuité de ces collaborations. Nous souhaitons montrer l'exemple, à l'instar de ce qui a déjà été fait au LPSC (laboratoire avec lequel les collaborations sont également fréquentes).

De plus ce projet permettra de nourrir la réflexion de l'éco-responsabilité pour la plate-forme dédiée à l'expérimentation que constitue Grid'5000 [5] et pour laquelle les membres des équipes MESCAL/MOAIS sont fortement impliqués. Ainsi plusieurs actions seront menées en accompagnement du projet pour recueillir un maximum d'informations par exemple sur la consommation, le recueil de trace d'activité afin d'évaluer la possibilité d'utiliser ce type de solution de free-cooling pour des grappes de Grid'5000. En prime les informations recueillies qui correspondront à l'exploitation d'une grille de production, seront mises à la disposition de la communauté Grid'5000. Ainsi nous renforcerons les liens entre grille de production et grille de recherche.

4 Fonctionnement

Le principe du freecooling est d'abord de bien confiner les zones froides (à l'avant des machines, là où elles aspirent l'air) et chaudes (à l'arrière, là où elles rejettent de l'air chaud).



Nous réalisons donc un capotage autour des 2 racks prévus (en gris sur le schéma) qui enveloppe complètement ces derniers. Un caisson de ventilation équipé d'un filtre aspire l'air de dehors et l'injecte en face avant des machines, dans la zone froide. L'air chaud est évacué par une ouverture vers l'extérieur à l'arrière des machines, dans la zone chaude, la différence de pression faisant sortir l'air naturellement à l'extérieur du bâtiment.

Une canalisation en zone chaude permet de reprendre de l'air chaud qui sera réinjecté si nécessaire vers l'entrée d'air du caisson afin de réguler la température, en particulier pour éviter les phénomènes de condensation. La régulation est contrôlée par un système d'asservissement électronique qui pilote les 2 registres (un pour l'entrée d'air principale et un autre vers le circuit de recyclage d'air chaud).

5 Réalisation

La circulation forcée de l'air est assurée par un caisson de ventilation intégré fournissant un débit de 4000m3/h, comprenant une entrée d'air, une sortie d'air, un moteur, un ventilateur, un filtre lavable et un pressostat de sécurité. Le pressostat permet de désactiver le système en cas d'encrassement trop important du filtre. Le caisson est suspendu au plafond à l'aide de tiges filetées et de patins anti-bruit.

Des gaines souples en aluminium de diamètre 300mm et 250mm permettent d'acheminer l'air entre les différents éléments. Elles sont reliées par des modules en galva (picages plats, raccords en T, registres) fixés au plafond à l'aide de kits de cablettes. Cela permet une grande modularité et rend le système facilement transposable dans d'autres situations. En outre, cela permet de faire facilement des modifications afin de compléter le dispositif. Par exemple, nous pourrons dérouter le soufflage de l'air froid en période hivernale afin de refroidir l'intégralité de la salle (pas seulement les racks) et arrêter complètement les clims qui sont utilisées pour les autres serveurs.

Le capotage est réalisé à l'aide de cornières perforées et de plaques de polycarbonate alvéolé de 16mm d'épaisseur. Les plaques sont fixées sur l'armature réalisée à l'aide des cornières. Une porte d'accès est aménagée en face avant dans la zone froide qui sert également d'accès éventuel aux consoles. La plaque arrière est facilement démontable afin de permettre l'accès à la zone chaude.

Les double vitrages des fenêtres ont été remplacés par une couche de plexiglas, une couche de bois et 2 couches d'isolant thermique. Il a ainsi été facile d'aménager les ouvertures et la fixation de grilles d'aération.

Les 2 registres permettant la régulation de température sont pilotés par des servo-moteurs industriels. Des sondes de températures sont installées dans les différentes zones. Un système autonome pilote les servo-moteurs, sondes de températures et vitesse du ventilateur

6 Gestion des ressources

La gestion des ressources se fait à 3 niveaux :

- Au niveau hardware, nous utilisons le protocole IPMI afin d'agir sur l'alimentation (on/off) des nœuds de calcul. Un système simple envoie des commandes d'arrêt lorsque la température devient trop élevée et signale au gestionnaire de ressources que les nœuds ne sont plus disponibles.
- Le gestionnaire de ressources OAR qui est utilisé pour le système de batch est informé de l'état de disponibilité des nœuds et gère les taches dans des files d'attente. Il est également capable d’arrêter lui-même des nœuds de calcul en cas d'absence de tâches à traiter afin d'économiser l'énergie.
- Le gestionnaire de la grille de calcul CIGRI communique avec OAR pour envoyer les tâches ou récupérer les résultats. Si des tâches ont été interrompues (par exemple des tâches qui étaient en cours au moment où les nœuds se sont arrêtés pour température trop importante), CiGri récupère

ces tâches et les envoie sur un autre calculateur disponible ou renverra ces tâches sur ce calculateur lorsqu'il sera à nouveau opérationnel.

7 Problèmes rencontrés

- Arrêt et reprise des jobs parallèles : Nous avons mis au point un mécanisme de reprise des jobs basé sur l'utilisation de la mise en veille des nœuds de calcul (suspend-to-ram ou suspend-to-disk). Le gestionnaire de ressources envoie un signal aux jobs avant l'extinction des nœuds pour synchroniser la suspension des processus parallèles (certaines librairies MPI prennent cela en charge par exemple) dans le but de pouvoir reprendre l'exécution des jobs là où ils en étaient lorsque la machine est à nouveau correctement refroidie. Cependant, nous nous heurtons actuellement à une limitation due à la gestion des cartes de réseau rapide qui ne reprennent pas correctement les communications à la remise en route. Le mécanisme fonctionne cependant pour les jobs qui n'utilisent pas le réseau rapide (jobs séquentiels ou à mémoire partagée), ce qui est le cas de la plupart des jobs « grille ».
- Impact sur la salle machine « hôte » : Nous avons fait le choix d'installer notre système dans une salle machine existante, qui est équipée d'une climatisation classique pour le refroidissement d'un certain nombre de serveurs déjà en place. Dans la théorie, notre système est indépendant et complètement isolé de cette salle. Ce choix a été fait pour des raisons pratiques de proximité physique et réseau. Cependant, l'inconvénient est que nos tuyauteries n'étant pas isolées, elles influent sur la température ambiante de la salle hôte et nous sommes parfois obligés d'arrêter notre machine, non pas parce que le refroidissement n'est plus possible, mais parce que nous réchauffons trop la salle hôte. Nous pensons réaliser un cloisonnement de la partie de la salle où notre dispositif est installé, avec du matériel isolant afin de réduire cet impact.

8 Bilan sur les 6 premiers mois de fonctionnement

Notre bilan sur cette première période estivale (de début avril à fin septembre) est très encourageant: nous avons mesuré un taux de disponibilité de 87%, ceci incluant une période d'arrêt total de 5 jours en aout que nous avons mis à profit pour améliorer l'étanchéité du dispositif et nettoyer les filtres. Le nombre d'interruptions est de 32 au total mais il tombe à 16 lorsque l'on enlève les 2 premiers mois de mise au point. Le principal problème rencontré est que nous ne sommes pas neutres vis-a-vis de la salle machines (climatisée) dans laquelle nous nous sommes installés (voir paragraphe ci-dessus). Il aurait été plus efficace de s'installer dans une pièce indépendante ou alors d'apporter plus d'importance à l'isolation (utiliser un caisson à double isolation, isoler les gaines).

La machine actuellement refroidie supporte en fait jusqu'à 35°C de température ambiante (classe 2 ASHRAE) , ce qui devrait nous permettre d'aller bien au delà des 85% de taux de disponibilité envisagé au départ (basé sur une température maxi de 26 degrés)

[1] https://ciment.ujf-grenoble.fr

[2] http://lpsc.in2p3.fr/informatique/ecoclim.html

[3] ASHRAE TC 9.9, 2011 Thermal Guidelines for Data Processing Environments Expanded Data Center Classes and Usage Guidance, http://tc99.ashraetcs.org/documents/ASHRAE%20Whitepaper%20-%202011%20Thermal%20Guidelines%20for%20Data%20Processing%20Environments.pdf

[4] http://www.liglab.fr/spip.php?page=mot&id_mot=187

[5] http://www.grid5000.fr