Retours d'expériences et perspectives sur les aspects énergétiques d'un mesocentre (Grenoble)

GreenDays@Luxembourg

Bruno Bzeznik, Pierre Neyron, Olivier Richard

CIMENT, LIG

28-29 Janvier 2013

Bruno Bzeznik

- Ingénieur de recherche sysadmin CIMENT
- Membre de l'équipe de développement OAR/CIGRI au LIG

Outline

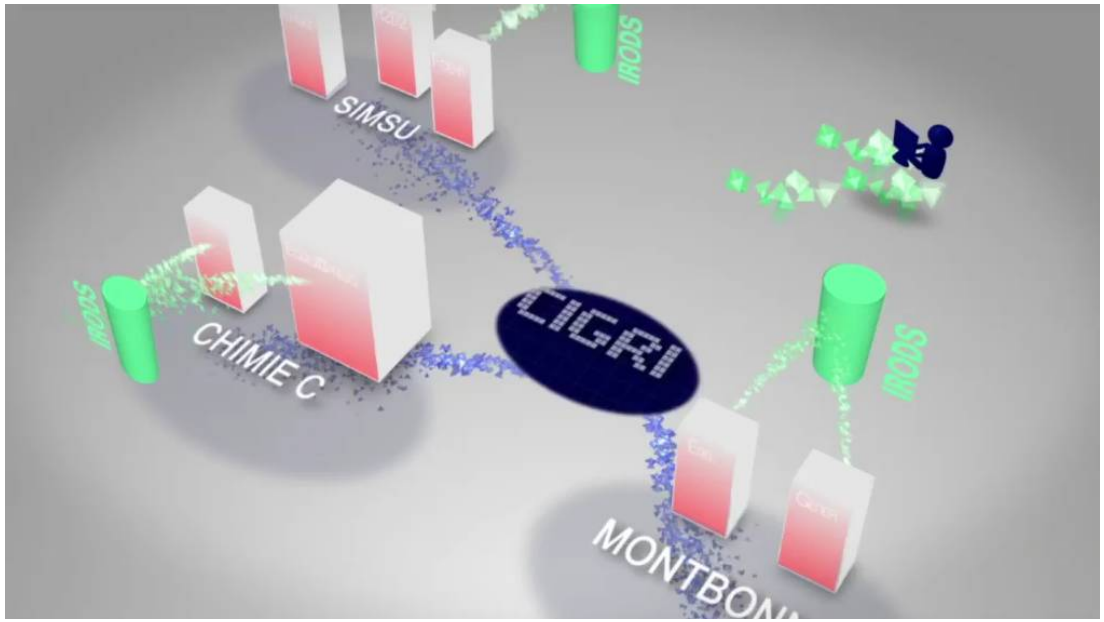
Outline

CIMENT

- CIMENT is the High Performance Computing (HPC) Centre of Grenoble University since 2000.
- It provides researchers and engineers with an easy access to local HPC resources to develop and test their codes
- It is composed of about 3500 cpu cores (2012, 5500 expected in 2013) in a dozen of supercomputers
- Collaborations with the LIG (Laboratoire d'Informatique de Grenoble) on batch scheduling problems.

CiGri

- CiGri is the grid middleware aggregating the computing power of the supercomputers
- Its goal is to optimize the usage of the (free) resources with regard to multi-parametric applications
- Cigri middleware developped by CIMENT/LIG (next release v3 in february)



Outline

Appels d'offre

- Dès 2005 : prise en compte de critères écologiques dans les appels d'offre
  - Efficacité énergétique des alims
  - Efficacité globale
  - Cycle de vie, recyclage, mtbf, etc...
  - Garantie 5 ans minimum...
- Critères placés au même niveau que le prix dans la pondération !

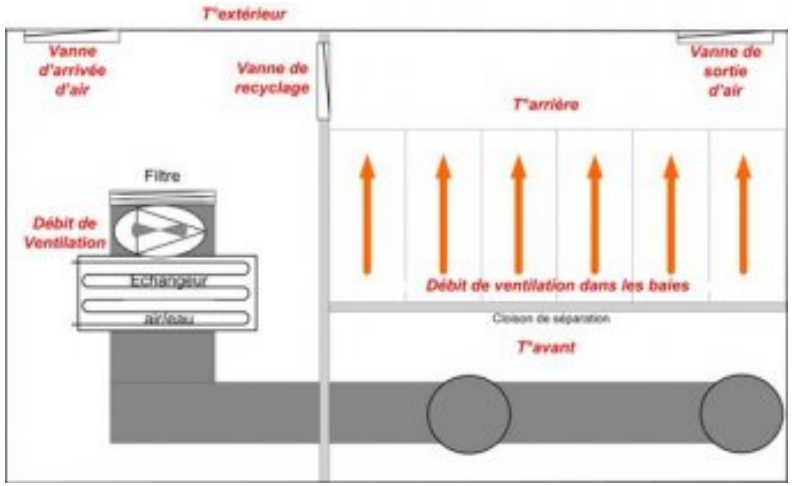
Vers une mutualisation des ressources de calcul

- A partir de 2007, CIMENT décide de regrouper les machines dans un seul datacenter, qui à défaut d'être efficace, permet au moins de faire ressortir les coûts de fonctionnement (prise de conscience)
- La mutualisation des machines est déja un premier pas vers plus d'efficacité énergétique

Ecoclim (credit : Bernard Boutherin)

En 2008, le LPSC, un laboratoire Grenoblois, met en place un datacenter en freecooling (Ecoclim)

- De nombreuses années de fonctionnement avec un bilan très positif : pour 80kW d'IT, c'est 160 000 kWh/an d'économisés.



Gestion de l'énergie dans OAR

- OAR est le gestionnaire de taches et de ressources qui exploite les machines HPC de CIMENT. Il est développé au sein du LIG (Laboratoire d'informatique de Grenoble).
- Dès 2006, il est capable de prendre en compte la puissance consommée par les noeuds de calcul dans son ordonnancement.
- Depuis 2009 il prend en charge l'arrêt et l'allumage automatique des noeuds de calcul en fonction de la charge.

Gestion de l'énergie dans OAR

OAR Cluster nodes

default summary

	Free	Busy	Total
network_address	4	15	32
resource_id	32	120	256

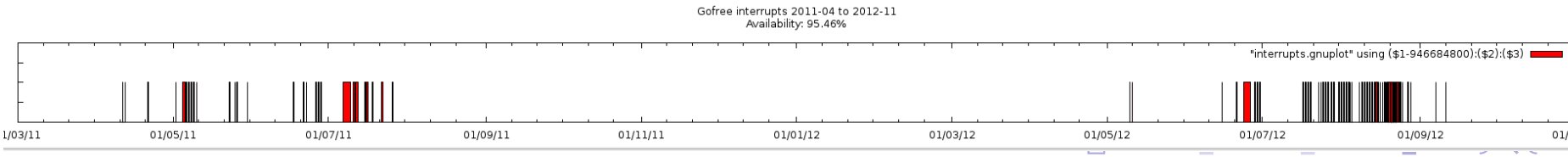
Reservations:

Reservations for property iru=0:

rli0n0	182270	182270	182270	182270	182270	182270	182270	182270
rli0n1	182270	182270	182270	182270	182270	182270	182270	182270
rli0n2	Free	Free	Free	Free	Free	Free	Free	Free
rli0n3	Free	Free	Free	Free	Free	Free	Free	Free
rli0n4	182267	182267	182267	182267	182267	182267	182267	182267
rli0n5	Free	Free	Free	Free	Free	Free	Free	Free
rli0n6	182271	182271	182271	182271	182271	182271	182271	182271
rli0n7	182271	182271	182271	182271	182271	182271	182271	182271
rli0n8	182271	182271	182271	182271	182271	182271	182271	182271
rli0n9	StandBy	StandBy	StandBy	StandBy	StandBy	StandBy	StandBy	StandBy
rli0n10	StandBy	StandBy	StandBy	StandBy	StandBy	StandBy	StandBy	StandBy
rli0n11	182294	182294	182294	182294	182294	182294	182294	182294

Frigid'r : freecooling extrême

- En 2011, tendance nette à l'augmentation de la température admise par les serveurs (35 voire 40 degrés pendant un certain % du temps)
- Réalisation d'un freecooling "extreme", c'est à dire sans clim ; fabrication "maison" pour un cout inférieur à 3000 euros autour d'un calculateur de 10kW.
- **Arrêt automatique de la machine en cas de fortes chaleurs**
- Bilan très positif :
  - **disponibilité de 96,4%** sur 2 ans de fonctionnement, PUE inférieur à 1.1.
  - Prise en compte de la contrainte par les chercheurs



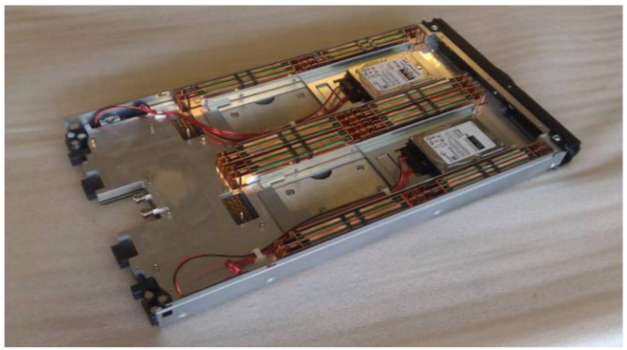
Equip@meso@Grenoble : un calculateur éco-responsable

- Budget d'un million d'euros (equipex)
- Dialogue compétitif (procédure longue : 1 an)
- Mise en production : **mai 2013**.
- Solution de refroidissement innovante (PUE de la partie calcul inférieur à 1.1) : **refroidissement à eau tiède 35/50°**

Equip@meso@Grenoble : un calculateur éco-responsable

- Chassis hydraulique

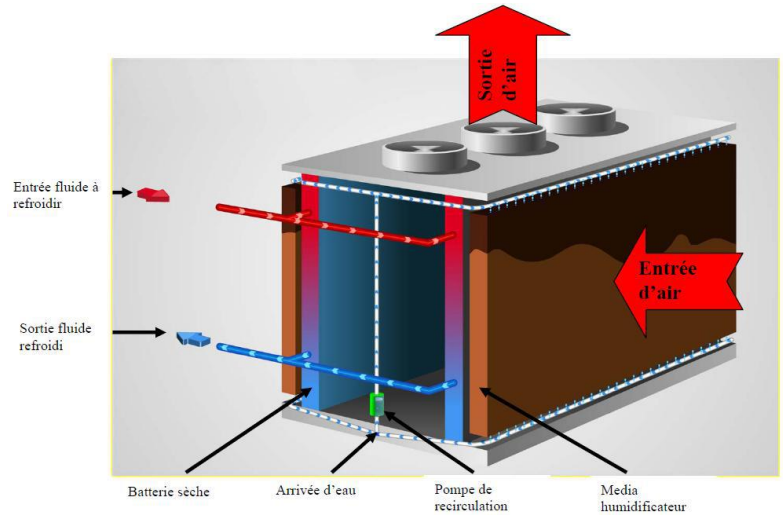
- Lame double serveur avec plaque de refroidissement hydraulique



Credit : Bull (C)

Equip@meso@Grenoble : un calculateur éco-responsable

- Tour de refroidissement hybride (ventilateur + échangeur air/eau + adiabatique)



Credit : Bull (C)

Outline

Outline

Serveurs haute densité : Clusters d'ARM/Atom/...

- L'idée est d'utiliser les processeurs des systèmes embarqués (smartphones,...) pour faire du HPC, car ces processeurs sont très optimisés au niveau de l'énergie
- Ex : RECS BOX Christmann (600 noeuds dans un rack !)
- Ex : SiCortex : Mips 64 bits : 5,832 cores pour 18 kW 5,8 téraflops en 2007 !
- Projet MontBlanc $\rightarrow$ Exascale avec des procs de l'embarqué (Arndaleboard)
- Pour le moment, surtout 32 bits, mais une guerre est lancée avec des nouveaux procs 64 bits (tegra3, Exynos 5, Mali T600,...)
- Puissances beaucoup plus faibles : **10-20**Kw/rack

Serveurs haute densité :Clusters d'ARM/Atom/...

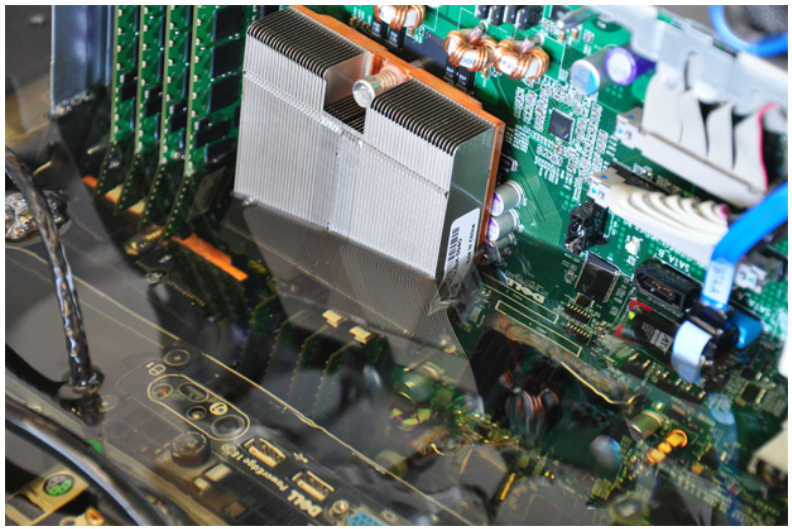
Outline

Direct Liquid Cooling

- Actuellement peu répandu (réservé au HPC)
- Sera certainement de plus en plus proposé par les constructeurs pour atteindre des densités élevées (80kW/rack)
- Inciter les constructeurs à suivre des standards !

Immersion

- Huile minérale
- Serveurs entiers, non modifiés (sans ventilos), plongés dans le bain
- Intel a validé le procédé sur un datacenter pendant un an. PUE=1.03 ! !
- Permetterai en plus d'augmenter la fréquence des processeurs



Immersion

Outline

Vers une vision intégrée

- Nécessaire à l'échelle d'un méso-centre
- OAR est ses modes/possibilité d'économie d'énergie (tunning et amélioration)
- ComputeMode pour l'usage des machines de bureaux
- Le monitoring orienté tâche et le monitoring de consommation d'énergie
- Détecter les applications énergétiquement inefficace.
- CiGri-v3 et OAR pour un ordonnancement global avec un critère de consommation
- Sensibilisation et information de l'utilisateur (Accounting : coût financier/conso des tâches, en place de métrique en tps réel pour l'aide à la mesure de l'intérêt des résultats et des tâches)

Cluster Virtuel - ComputeMode

- Création d'un cluster virtuel avec les ressources inutilisées
- Exemple salle de TP la nuit (UFRIMA - Université Joseph Fourier)
  - PXE
  - Wake-On-Lan
  - Diskless systems
  - OAR comme gestionnaire de ressources, **réveil à la demande, zone indisponible**
- **Usage : cluster d'appoint intégré dans la grille du Méso-centre CIMENT**
- **Heure creuse, pas de climatisation, disques inutilisés ! :)**

Vers le monitoring de (ressources par) tâche

Monitoring d'usage global des ressources par node : type Ganglia

- Donne une vue globale d'usage de la plateforme
- Difficile de faire le lien avec les tâches (surtout en analyse post-mortem)

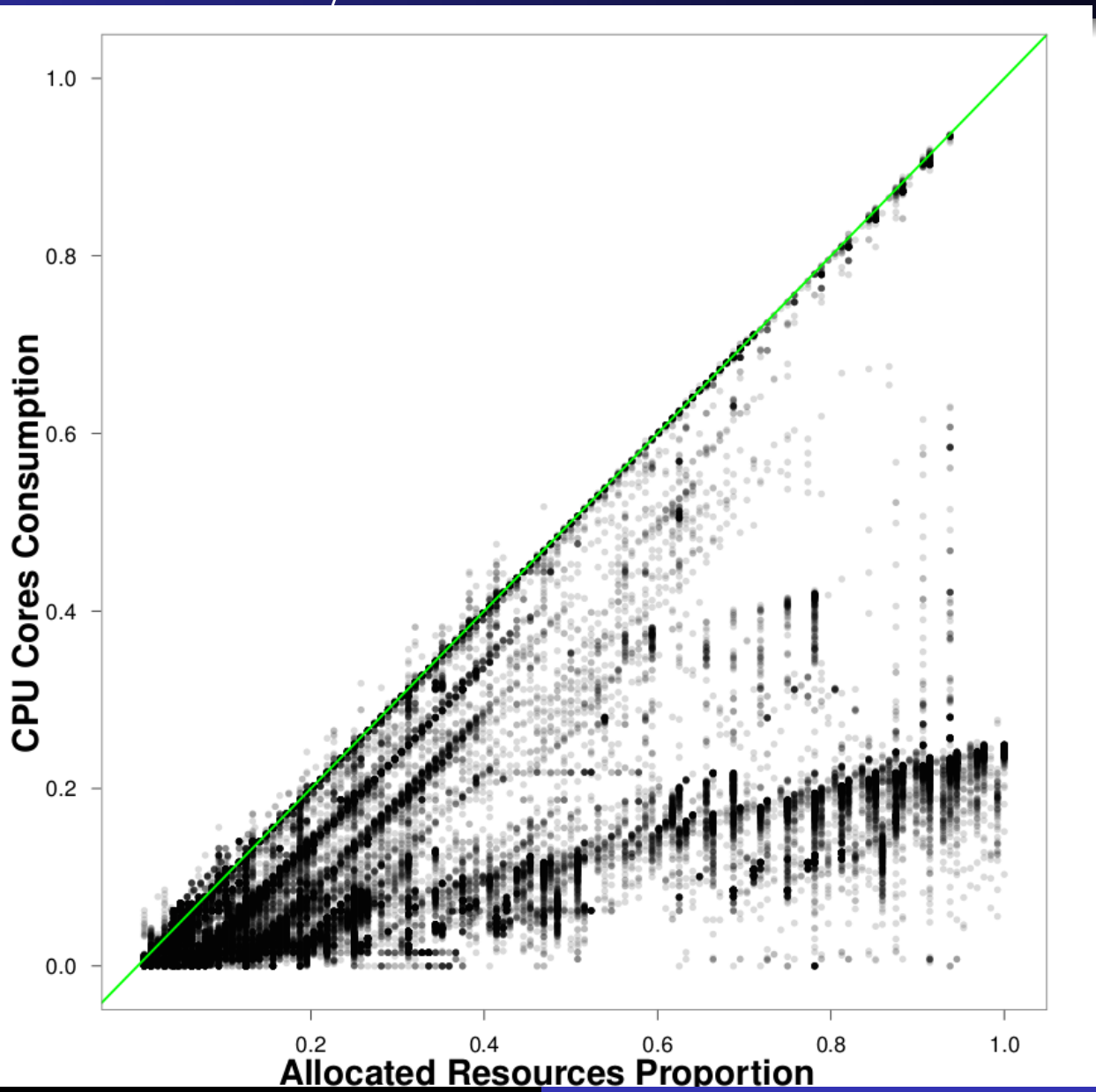
Suivi de conso **SandyBridge**

Intel : **RAPL** (running average power limit) provides MSRs reporting the total amount of energy consumed (**updated every 1msec**) by the **package/core/uncore/dram**

Monitoring de tâches

- Tâche confinée via les CPUSET (monitoring par CPUSET).
- Freq. échantillonnage : 1 min (surcoût / stockage).
- Projet **Colmet** disponible **au printemps 2013**.

Consommation CPU / ressources CPU allouées



Outline

Conclusion

Etat des lieux

- Préoccupation qui s'est accentuée sur une dizaine année
- **Facturation de la consommation est maintenant sur le même plan que l'achat de matériel**
- ...

Perspectives

- Suivre l'évolution matérielle (l'arrivée des solutions de l'embarqué avec GPGPU dans les serveurs ? ? ?)
- **Vision intégrée de la consommation de l'énergie à l'échelle du Méso-centre nécessaire pour une m eilleur optimisation**

Merci !

- Questions