Techniques de limitation de la consommation énergétique d'un nœud de calcul

Green Days 2024

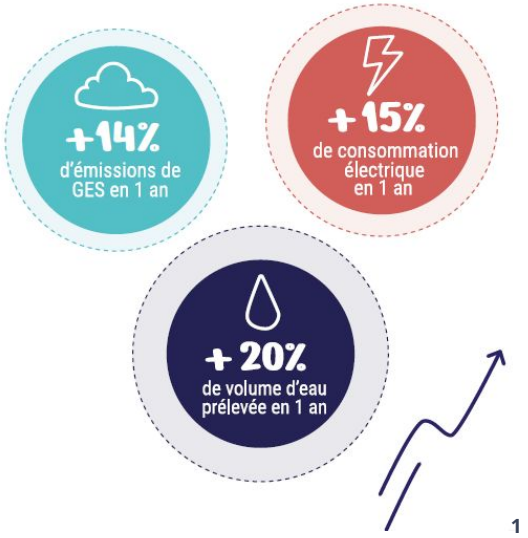
Vladimir Ostapenco [1]
Laurent Lefevre [1]
Anne-Cécile Orgerie [2]
Benjamin Fichel [3]

[1] Univ Lyon, EnsL, UCBL, CNRS, Inria, LIP, Avalon
[2] Univ Rennes, Inria, CNRS, IRISA, Magellan
[3] OVHcloud

FrugalCloud (Défi Inria/OVHcloud)

Contexte

- Augmentation considérable de l'utilisation des technologies de l'information et de la communication
- Centres de données et services cloud
  - Sont au centre de cette croissance
  - Leur nombre et leurs impacts sont en constante augmentation
- **De 2021 à 2022 en France**, les opérateurs de centres de données [1]
  - **+14 %** d'émissions de GES
  - **+15 %** de consommation électrique
    - **Plus de 90 % des émissions** globales de GES



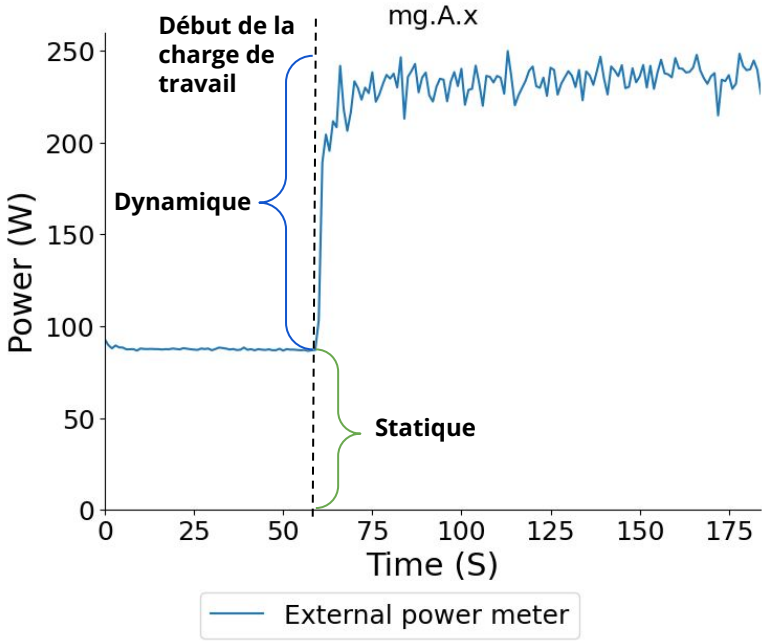
[1] ARCEP (2024). "Pour un numérique soutenable"

Contexte

- Dans ce contexte, l’efficacité énergétique des datacenters et des services cloud est une préoccupation croissante
- Nœuds de calcul représentent une part importante de la consommation totale des datacenters
- Nous devons avoir les moyens de
  - Mesurer la consommation d'énergie
    - Comprendre les impacts
  - **Limiter la consommation énergétique**
    - Réduire les impacts
    - Améliorer l’efficacité énergétique

Consommation d’un nœud de calcul

- **Partie Statique**
  - Consommation quand aucune charge de travail n'est exécutée
  - Représente une partie non négligeable de la consommation d’un nœud de calcul
- **Partie Dynamique**
  - Augmentation de la consommation due à l'exécution de la charge de travail



Profil de puissance avant et pendant l'exécution du benchmark MG NAS

Techniques de limitation de la consommation énergétique

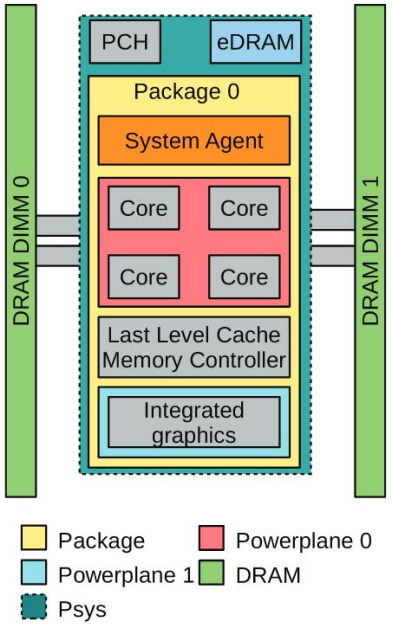
- **Shutdown policies** (agit sur la consommation statique)
  - Arrêt et allumage des noeuds de calcul
- **Sleep states** (agit sur la consommation statique)
  - Mise en veille ou allumage/extinction des ressources
- **DVFS** (agit sur la consommation dynamique)
  - Modulation de tension et de fréquence du CPU
  - Disponible sur la majorité des nœuds de calcul
- **Intel RAPL** (agit sur la consommation dynamique)
  - Limitation de puissance des composants du nœud de calcul (package CPU et DRAM)

Application de chaque technique engendre des coûts en termes de temps et d'énergie non négligeables car la transition d'état nécessite du temps et implique une consommation d'énergie. [1,2]

[1] Rais et al. (2018). Quantifying the impact of shutdown techniques for energy-efficient data centers
[2] Kim et al. (2008). System level analysis of fast, per-core DVFS using on-chip switching regulators

Intel RAPL - Running Average Power Limit

- Introduit dans l'architecture Intel Sandy Bridge (2011)
- Bien que la plupart des travaux étudient le RAPL pour la mesure [1,4], **conçu** à l'origine **à des fins de limitation de puissance**
- Permet **la spécification de la puissance moyenne sur une période de temps** sur les domaines de puissance
- Assez précis et stable pour la majorité des applications de longue durée [2]
- A des temps de transition d'état courts et est assez efficace [2]
- Tend à remplacer DVFS [3]
- **Pouvons-nous utiliser RAPL pour un plafonnement strict de la puissance** afin de répondre aux contraintes énergétiques ?



Domaines de puissance disponibles dans RAPL [3]

[1] Jay et al. "An experimental comparison of software-based power meters: focus on CPU and GPU."
[2] Zhang et al. (2015). A quantitative evaluation of the RAPL power control system
[3] Imes et al. (2019). CoPPer: Soft Real-Time Application Performance Using Hardware Power Capping
[4] Khan et al. (2018). RAPL in Action: Experiences in Using RAPL for Power Measurements

Intel RAPL - Comment RAPL fonctionne ?

- Il n'existe que **peu d'information officielle publiquement accessible** sur la mise en œuvre interne du RAPL
  - Travail relativement ancien décrivant un mécanisme d'estimation et de limitation de la puissance de la mémoire vive [1]
  - Seule l'interface de configuration et d'utilisation est documentée
- **Aucune information officielle** sur
  - Propriétés telles que la précision, le temps de stabilisation et les impacts sur les performances des applications
  - Différences d'implémentation entre les architectures de processeur
  - Comment la réduction de puissance est obtenue
- Informations **trouvées dans la littérature**
  - RAPL utilise le DVFS et d'autres techniques qui forcent les composants du processeur à rester inactifs à de faibles niveaux de puissance [2]
  - UFS (Uncore Frequency Scaling) semble être utilisé par le mécanisme de contrôle de puissance Intel RAPL[3]

[1] David et al. (2010). RAPL: memory power estimation and capping
[2] Zhang et al. (2016). Maximizing Performance Under a Power Cap: A Comparison of Hardware, Software, and Hybrid Techniques
[3] Riha et al. (2020). Evaluation of DVFS and Uncore Frequency Tuning Under Power Capping on Intel Broadwell Architecture

Intel RAPL - Analyse de la technologie RAPL

- **Validation du mécanisme de limitation de la puissance**
  - En utilisant un wattmètre externe de haute précision
  - Pour les charges de travail hétérogènes
- **Étude des caractéristiques**
  - Valeurs de limitation de puissance prises en charge
  - Précision
  - Granularité minimale de limitation
  - Temps de stabilisation
- **Étude des mécanismes utilisés** pour ajuster la puissance
- **Étude des impacts sur les performances** des applications
  - FLOPS et bande passante mémoire

Pour les benchmarks

- **NAS** (EP, MG, LU)
- **Stream**

Sur **4 clusters de Grid'5000**

- **Taurus** (Sandy Bridge - 2011 - ***RAPL v1***)
- **Orion** (Sandy Bridge - 2011 - ***RAPL v1***)
- **Gemini** (Broadwell - 2014 - ***RAPL v2***)
- **Troll** (Cascade Lake - 2019 - ***RAPL v2***)

Intel RAPL - Validation du mécanisme RAPL

Quelques observations

- RAPL est capable de faire respecter les limites de puissance avec **un bon niveau de précision**
- **Puissance** du noeud de calcul **est effectivement réduite**
- **Dépassements** de la limite de puissance **sont possibles** et sont fréquentes pour les charges de travail instables (profil de consommation instable)

Intel RAPL - Étude des caractéristiques

- **Précision**
  - Calcul de MAPE (Mean Average Percentage Error) entre la limite imposé et la puissance donnée par Intel RAPL
  - Valeurs **MAPE** sont presque toujours **inférieures à 2 %**
  - MAPE inférieur à 0.1 % pour les benchmarks NAS EP et Stream sur les noeuds avec Intel RAPL v2
- **Granularité minimale de limitation**
  - Limitation de puissance est exprimée en “Power Units” (par défaut 0.125 Watts)
  - Vérifier si le pas minimum est de 0,125 Watts et si la puissance peut réellement être affectée par ces pas
  - Puissance **peut être influencée par des incréments de 0,125 Watts** pour chaque cluster étudié
- **Temps de stabilisation**
  - Temps nécessaire pour appliquer une limite de puissance et stabiliser la consommation
  - Resultats: *RAPL v1* - **plus de 5 secondes**, *RAPL v2* - **entre 0.5 et 1 seconde**

Intel RAPL - Étude des mécanismes utilisés

Comment Intel RAPL gère la fréquence CPU core et uncore pour limiter la puissance ?

- **Fréquence core est significativement affecté** par les limitations de puissance (de la manière équivalente à la puissance)
  - Corrélation entre la fréquence core et les limitations de puissance est très élevé
- **Fréquence uncore est également impacté** par les limitations, mais la dépendance avec la puissance est moindre
- Gestion des fréquences dépend de la nature de la workload exécuté
  - Différente pour les workloads CPU et RAM-intensives

Intel RAPL - Étude des impacts sur les performances

Quel est l'impact de la limitation avec Intel RAPL sur les performances des applications en termes de **FLOPS** et de **bande passante mémoire** ?

- **Valeur FLOPS diminue** de la même manière que la fréquence core du CPU pour le benchmark NAS EP (CPU-intensive)
- **Bande passante mémoire est moins impacté** par l'application de limitation de puissance
- Existence de la plage de limites de puissance où **la puissance peut être réduite sans modifier la bande passante mémoire** pour les benchmarks Memory-intensive
    - RAPL est donc capable de réduire la puissance sans dégrader les performances

Intel RAPL - Conclusions

- **Nous avons validé que RAPL**
  - Capable de faire respecter les limites de puissance avec un bon niveau de précision, en particulier pour les charges de travail intrinsèquement stables (profil de consommation stable)
  - Est assez précis et a des faibles temps de stabilisation
  - Capable influencer la puissance par les incréments de **0.125 Watts**
  - Utilise **DVFS**, **UFS** et impose l'**état inactif de certains composants** internes du processeur à de faibles niveaux de puissance pour limiter la consommation d'énergie
- **Nous avons decouvert que**
  - Disponibilité des domaines de puissance RAPL utilisables pour la limitation dépend non seulement de l'architecture du processeur, mais également du BIOS et de son firmware
  - **Ne doit pas être considéré comme un levier de plafonnement strict de la puissance**, car la limite de puissance spécifiée peut être dépassée
  - Est capable dans certains cas de réduire la consommation sans dégrader les performances
  - Modulation des fréquences core/uncore effectuée dépend de la nature du workload
- **Future Works**
  - Publier les travaux sur Intel RAPL (le travail est tout frais, ne le dévoilez pas)

Merci pour votre attention!

Avez-vous des questions ?