

Monitoring énergétique dans Slices-FR/Grid'5000 et le logiciel Kwollect

Simon Delamare

LIP / CNRS

Section 1

Introduction

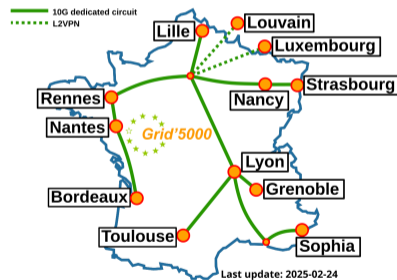
Introduction

- Présentation
- Retour d'expérience sur le monitoring énergétique dans la plateforme Slices-FR/Grid'5000 et sur le logiciel Kwollect
 - ▶ Historique du monitoring sur Slices-FR/Grid'5000
 - ▶ Exemples de travaux de recherche
 - ▶ Présentation de Kwollect
 - ▶ Illustration de l'utilisation de Kwollect
- Remerciements

L'infrastructure de recherche Slices-FR/Grid'5000

Réalisation d'expériences dans tous les domaines de l'informatique distribuée :
calcul HPC/IA, virtualisation, réseau, etc.

- ~500 nœuds réservables
 - ▶ entièrement reconfigurables (*root*, déploiement d'OS, ...)
 - ▶ 10000 cores / 200 GPUs
 - ▶ Majoritairement de type HPC, beaucoup de diversité matérielle
 - ▶ Répartie sur 10 sites géographiques



L'infrastructure de recherche Slices-FR/Grid'5000 (2)

- Pour la communauté de la recherche en informatique
 - ▶ 500+ utilisateurs actifs et 100-150 publications / an
 - ▶ Soutenu par les principaux instituts de l'ESR
 - ▶ Organisé en GIS, labellisé IR
 - ▶ Existe depuis plus de 20 ans !



L'infrastructure de recherche Slices-FR/Grid'5000 (2)

- Pour la communauté de la recherche en informatique
 - ▶ 500+ utilisateurs actifs et 100-150 publications / an
 - ▶ Soutenu par les principaux instituts de l'ESR
 - ▶ Organisé en GIS, labellisé IR
 - ▶ Existe depuis plus de 20 ans !



L'infrastructure de recherche Slices-FR/Grid'5000 (2)

- Pour la communauté de la recherche en informatique
 - ▶ 500+ utilisateurs actifs et 100-150 publications / an
 - ▶ Soutenu par les principaux instituts de l'ESR
 - ▶ Organisé en GIS, labellisé IR
 - ▶ Existe depuis plus de 20 ans !



Permet d'étudier l'énergie des infra informatiques depuis de nombreuses années.

- Possibilité d'être *root* → Modification des paramètres d'économie d'énergie, etc.
- Mesure des consommations à la prise + Mise à disposition des données aux utilisateurs

Section 2

Historique du monitoring énergétique dans Grid'5000

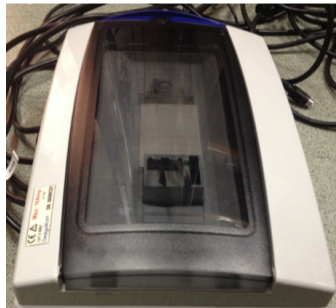
2007: Les débuts

- Premières discussions sur l'efficacité énergétique
- Démarrage des activités :
 - ▶ Projet Inria ARC Green-Net « Power aware software frameworks for high performance data transport and computing in large scale distributed systems » à Lyon, Toulouse, Grenoble et Virginia Tech
 - ▶ 1er challenge : comment monitorer ??

2007: Les débuts

- Premières discussions sur l'efficacité énergétique
- Démarrage des activités :
 - ▶ Projet Inria ARC Green-Net « Power aware software frameworks for high performance data transport and computing in large scale distributed systems » à Lyon, Toulouse, Grenoble et Virginia Tech
 - ▶ 1er challenge : comment monitorer ??

Wattmetre (version 0) par l'entreprise
Omegawatt (aujourd'hui *ADECWatts*)



2008: Déploiement et premiers prototypes

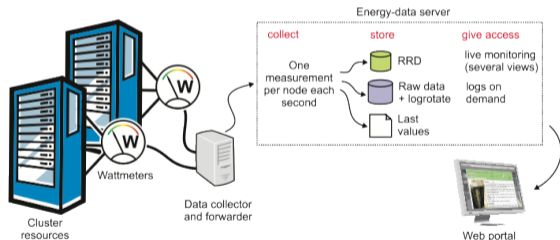
- Déploiement de wattmetres à Grid'5000 Lyon, Toulouse, Grenoble
- Prototype « Green Grid5000 » et démonstration au Inria Booth à SuperComputing 2008

2008: Déploiement et premiers prototypes

- Déploiement de wattmetres à Grid'5000 Lyon, Toulouse, Grenoble
- Prototype « Green Grid5000 » et démonstration au Inria Booth à SuperComputing 2008

Outil Grid'5000 : *Green Grid'5000*

- Stockage long terme RRD
- Pas d'API mais téléchargement des valeurs brutes via page PHP



2010: Large échelle

- Le site de Grid5000 Lyon complètement monitoré
 - ▶ Wattmetres version 1 : 150+ prises mesurées, 1 mesure par seconde
- Projet Inria Hemera pour l'expérimentation large échelle

2010: Large échelle

- Le site de Grid5000 Lyon complètement monitoré
 - ▶ Wattmetres version 1 : 150+ prises mesurées, 1 mesure par seconde
- Projet Inria Hemera pour l'expérimentation large échelle

Outil Grid'5000 : *Kwapi*

- API et interface graphique
- Support initial des métriques réseau
- Stockage HDF5



2010s : Démocratisation du monitoring

- Installation de PDUs monitorables par prise
- Consommations remontées par les interfaces d'administration (BMC)
- Monitoring logiciel avec RAPL

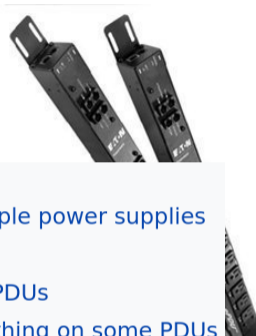


2010s : Démocratisation du monitoring

- Installation de PDUs monitorables par prise
- Consommations remontées par les interfaces d'administration (BMC)
- Monitoring logiciel

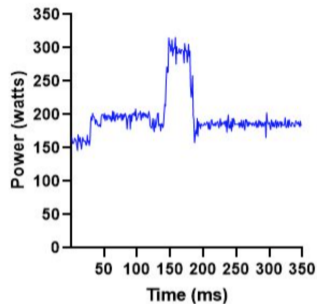
measurement artifacts and pitfalls

- 3.1 Kwapi handling of nodes with multiple power supplies
- 3.2 Measurement thresholds on PDUs
- 3.3 Noise measurement on some APC PDUs
- 3.4 Measurement resolution and smoothing on some PDUs
- 3.5 checking and calibration of the measurements
 - 3.5.1 First experiment (Zimmer at 20Hz)
 - 3.5.2 Second experiment (Zimmer at 10Hz)



2020s : Monitoring avancé

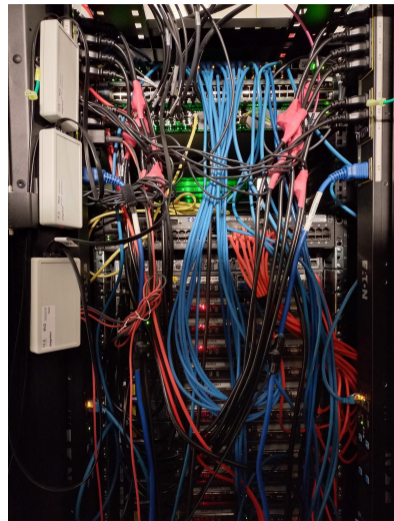
- 2020: wattmetres version 2 :
 - ▶ 50 mesures par seconde
 - ▶ déployés sur Grid'5000 Lyon, Grenoble, Nancy



© INRIA / Université de Lyon

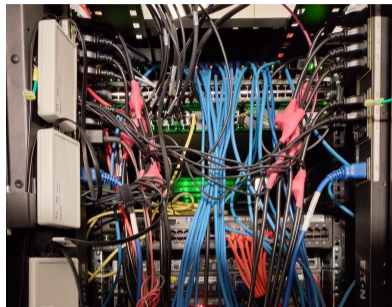
2020s : Monitoring avancé

- 2020: wattmetres version 2 :
 - ▶ 50 mesures par seconde
 - ▶ déployés sur Grid'5000 Lyon, Grenoble, Nancy
- 2022: wattmetres version 3 :
 - ▶ déploiement (en cours) : Lyon, Rennes, Lille, Strasbourg... l'ensemble des noeuds Slices-FR/Grid'5000



2020s : Monitoring avancé

- 2020: wattmetres version 2 :
 - ▶ 50 mesures par seconde
 - ▶ déployés sur Grid'5000 Lyon, Grenoble, Nancy
- 2022: wattmetres version 3 :
 - ▶ déploiement (en cours) : Lyon, Rennes, Lille, Strasbourg... l'ensemble des noeuds Slices-FR/Grid'5000



Outil Grid'5000 : Kwapi → *Kwollect*

- Support de diverses sources (pas que énergie)
- Amélioration de l'API
- Meilleur passage à l'échelle
- Nombreuses nouvelles fonctionnalités : Intégration avec le job scheduler, métriques utilisateurs, etc.

Section 3

Exemples de travaux de recherche

Exemples de travaux de recherche

Efficacité de l'utilisation des leviers (DVFS, Extinction, Power capping, ...)

- *Even On/Off is much more complex than expected !*
- *GreenFactory an orchestration framework
Combination: CVV + #MPI/#OpenMP Leverages*
- *Energy-efficient Resource Provisioning for Cloud Databases*
- *Going slower to go faster !*
- *Study the impact of CPU-Frequencies Scaling in Data-intensive clusters*

Nouvelles architectures

- *Energy proportionality on hybrid architectures*
- *Thermal and Power monitoring: The SeDuCe testbed*

Modélisation de la consommation

- *Understanding the energy consumption of blockchains: a focus on smart contracts*
- *Simulating Energy Consumption of MPI applications*
- *Monitoring energy consumption during training*
- *From hardware sensors to software energy sensors*

Consommation des datacenter / Cloud

- *Including IaaS users with an energy/performance VM lever*
- *Evaluating edge clouds' energy consumption*
- *Scheduling live-migration for fast and energy-efficient reconfiguration*

Even On/Off is much more complex than expected !

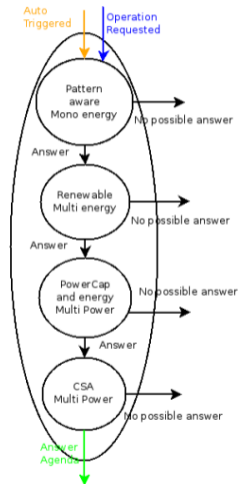
G5K@Lyon&Rennes

Pattern, power capping and energy aware : iif the overall nodes under supervision are respecting the fixed constraints

Electric Provider aware : iif the modular constraints imposed by the electric provider are respected

Cooling aware : iif the recommended thermal envelop is respected

Switching On & Off resources at ultra lage scale can take a loooooong time : minutes, hours,.....



Understanding the energy consumption of blockchains: a focus on smart contracts

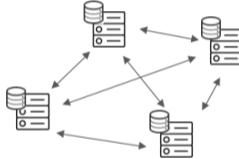
Expérimentations *in situ*

Analyses de performances de réseaux privés Ethereum & Hyperledger
 -Déploiements sur G5K de 2 à ~40 noeuds
 -Montée en charge
 -Déploiement et exécution de smart contracts



Contributions clés

- Développement d'un framework pour le benchmark de blockchains
- Modélisation du coût d'exécution de smart contracts sur Ethereum
- Illustration de l'impact négatif de la taille grandissante sur leurs performances



Mieux comprendre le coût énergétique des *blockchains* et de leurs *smart contracts*

Analyse du trafic existant

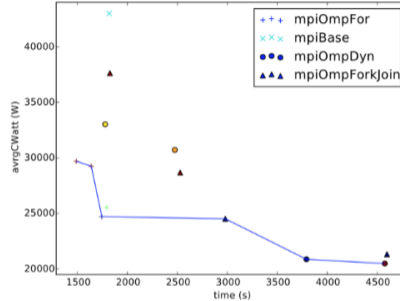
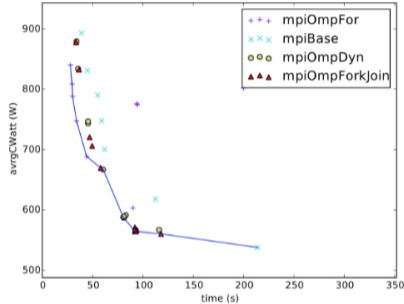
Déploiement & synchronisation d'un noeud sur le réseau public Ethereum sur G5K
 -Analyse des données publiques Ethereum



Contributions clés

- Estimation du nombre de contrats déployés mais non utilisés
- Modélisation de l'impact de ces contrats non utilisés sur les performances des blockchains
- Proposition d'un protocole de suppression des contrats non utilisés

GreenFactory an orchestration framework Combination: CVV + #MPI/#OpenMP Leverages



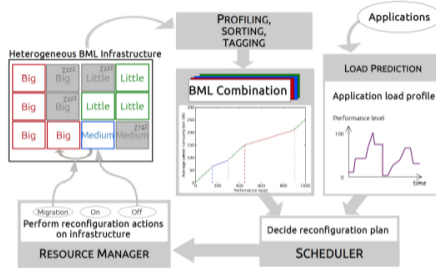
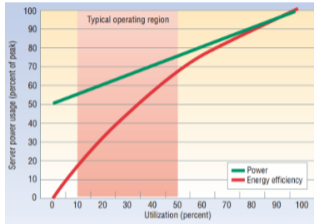
(K. A) Grid5000 (4 nodes – 48 cores) - (K. B) CURIE machine (128 nodes- 2048 cores)

How to orchestrate the usage of energy leverages available through time on moduable IT resources to answer fixed constraints ?

Energy proportionality on hybrid architectures

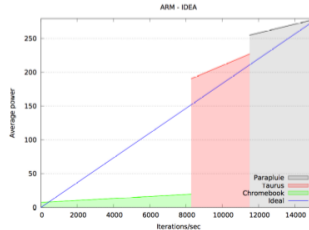
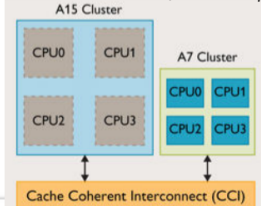


- Hybrid computing architectures : low power processors, co processors, GPUs...
- Supporting a "Big, Medium, Little" approach : the right processor at the right time



« The case for Energy-Proportional Computing »

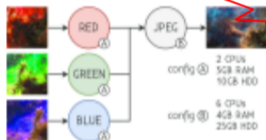
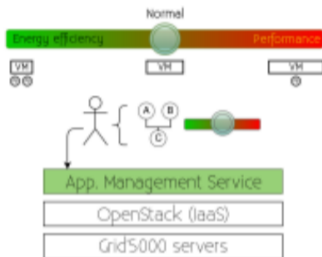
L. A. Barroso and U. Hölzle, *IEEE Computer*, 2007



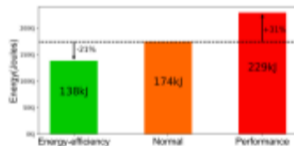
Violaine Villebonnet,
Georges Da Costa,
Laurent Lefevre,
Jean-Marc Pierson
and Patricia Stolf.
"Big, Medium, Little" :
Reaching Energy
Proportionality with
Heterogeneous
Computing
Scheduler", *Parallel
Processing Letters*,
September 2015

Including IaaS users with an energy/performance VM lever

David Guyon, Anne-Cécile Orgerie and Christine Morin



G5K@Rennes



- ▶ users select an execution mode
- ▶ impact on size of VMs
- ▶ bin packing algo places the VMs
- ▶ multiple deployment of *Montage* on Grid'5000

- ▶ loss in performance helps to consolidate the resources
- ▶ up to 21% energy saving in comparison with normal mode

Monitoring energy consumption during training



Goal: Compare energy consumption at training on different sites

- Get an incentive (not perfect estimate)
- Simple process (using CodeCarbon)

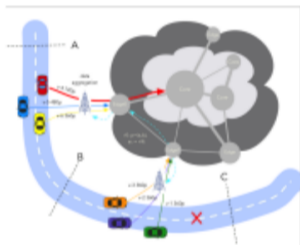
How: Benchmark the training of a small network (~1M parameters) on G5k

Large Impact	Minor impact
GPU used	GPU load (tested through batch size)
Additional take-away	
50%+ energy used to achieve the last 5 % increase in performance...	

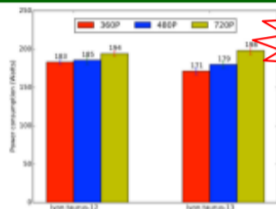
Evaluating edge clouds' energy consumption

Yunbo Li, Anne-Cécile Orgerie, Ivan Rodero, Manish Parashar and Jean-Marc Menaud

- ▶ Data stream analysis from cameras embedded on vehicles



- ▶ How to decide to compute at the edge or offload at the edge depending on QoS and energy-efficiency for a given IoT application?



G5K@Rennes
&Nantes

Power consumption for 4 small VMs Taurus-12 and 1 large VM on Taurus-13 for same amount of computation

- ▶ Higher power consumption with one large VM, but more energy-efficient
- ▶ Depends on the applications' resource requirements

« Scheduling live-migrations for fast and energy-efficient reconfiguration »

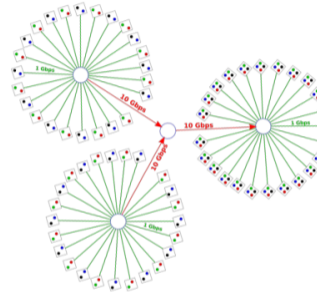


Modeling the VM migration protocol wrt. the network and the memory activity to wiser migration plans.

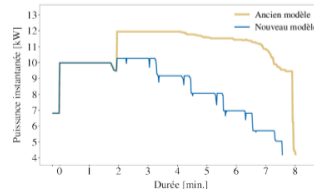
32.45 yrs/CPU of experiments to understand the problem and evaluate the solutions.

Biggest experiment: decommissioning 48 servers to 24 new ones.

21.5% less energy than state-of-the-art approaches



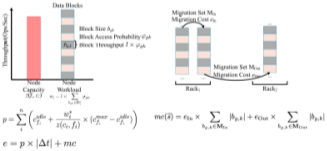
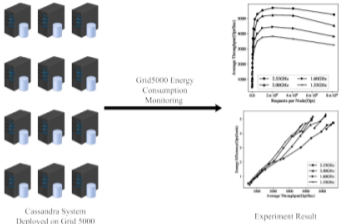
Vincent Kherbache, Eric Madelaine, Fabien Hermenier. Scheduling Live Migration of Virtual Ma-chines. IEEE transactions on cloud computing, 2017,





Energy-efficient Resource Provisioning for Cloud Databases

- Research Achievements
 - Energy efficiency of cloud databases under DVFS technique is analyzed.
 - A generic model is proposed for cloud systems. Then the model is specialized to improve resource provisioning within cloud database systems.
 - A frequency selection bounded problem is solved by nonlinear programming approach and multi-phases algorithm.
 - A frequency selection optimization problem is solved by genetic algorithm and monte carlo tree search algorithm.
 - A migration approach is proposed to further improve the energy efficiency of the system.
- Experiment with Grid5000
 - Cassandra. 10 nodes from Nancy Site;
 - YCSB benchmark;
 - DVFS technique: 8 frequency options;
 - Energy Consumption Monitoring with Kwapi;
 - 12 query workloads.

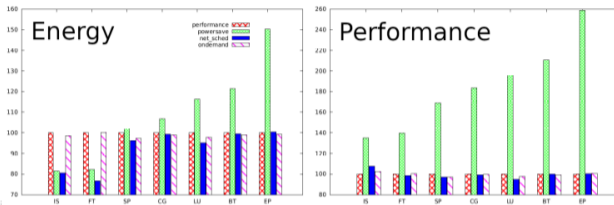


Going slower to go faster !



DVFS for HPC

- Difficult as processors are fully loaded
- Our proposal **net_sched** reduces frequency when not cpu-bound
- Decisions every 100ms
- Previously based on thresholds, now investigation on Machine Learning methods



Examples:

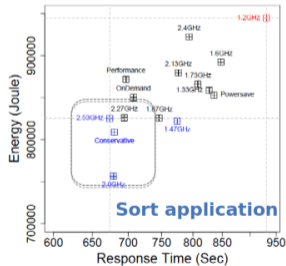
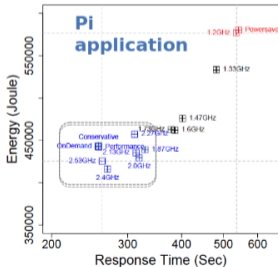
- FT bench: 23% less energy, 2% faster
- LU bench: 5% less energy, 5% faster

Going slower reduces the temperature
More available thermal capacity when it matters !

Study the impact of CPU-Frequencies Scaling in Data-intensive clusters

MapReduce applications show significant variation in CPU load during their running time

➤ Potential for saving energy by using **DVFS**



Different DVFS settings are not only sub-optimal for different MapReduce applications but also sub-optimal for different stages of the MapReduce application

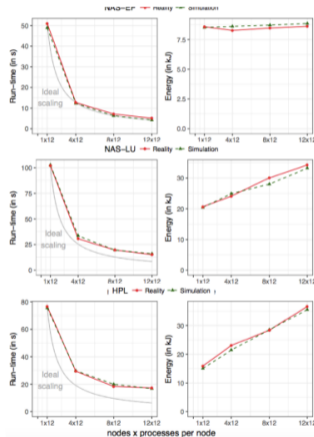
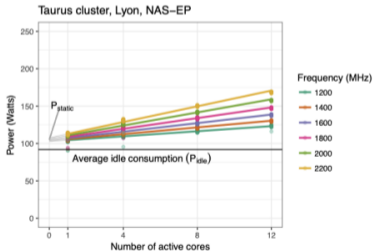
Governing energy consumption in Hadoop through CPU frequency scaling: An analysis

Shadi Ibrahim, Tien-Dat Phan, Alexandra Carpen-Amarie, Houssein-Eddine Chihoub, Diana Moise, Gabriel Antoniu *FGCS Journal 2016*

Simulating Energy Consumption of MPI applications

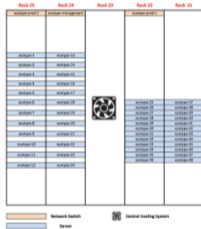
Univ. of
Grenoble

- ▶ Hac Specis project
- ▶ Energy models of multicore CPUs in SimGrid
- ▶ Validated against real measurements

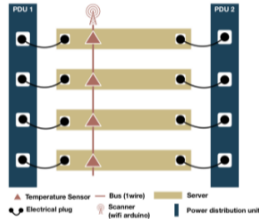


F. C. Heinrich *et al.*, "Predicting the Energy-Consumption of MPI Applications at Scale Using Only a Single Node," *2017 IEEE International Conference on Cluster Computing (CLUSTER)*, 2017,

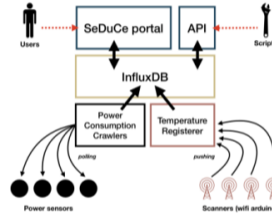
Thermal and Power monitoring: The SeDuCe testbed



Near 1000 Threads,
6To of RAM,
20To of SSD

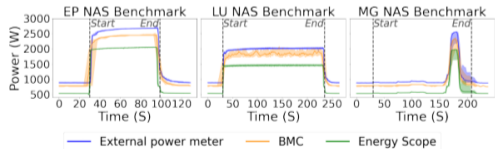
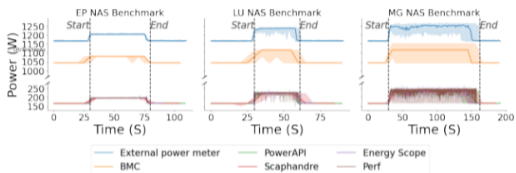


Power (PDU) and
thermal (specific)
monitoring,
every seconds



Data available through
an open portal and API

From hardware sensors to software energy sensors



Breakend/experiment-impact-tracker

ML CO2 Impact



24

Energy scope

« FrugalCloud » Inria&OVHcloud

Mathilde Jay, Vladimir Ostapenco, Laurent Lefevre, Denis Trystram, Anne-Cécile Orgerie and Benjamin Fichel. "An experimental comparison of software-based power meters: focus on CPU and GPU", CCGrid 2023 - 23rd IEEE/ACM international symposium on cluster, cloud and internet computing, Bangalore, India, May 1-4, 2023

Section 4

Le logiciel Kwollect

Motivation

- Monitoring d'infrastructure, focus sur l'environnemental :
 - ▶ Consommation électrique à la prise
 - ▶ Température ambiante
 - ▶ Métrique des BMC, des PDUs
 - ▶ ...
- Grand volume, haute fréquence (*Wattmetre* à 50 mesures/s)
- Conservation longue durée et sans perte possible
- Orienté vers les utilisateurs
 - ▶ Pas un outil pour les admin de la plateforme
 - ▶ API pour récupérer les données, intégration avec le job scheduler.
 - ▶ Métriques "custom"



- Impossible de laisser l'utilisateur se débrouiller seul
- Outils de monitoring traditionnels non adaptés (*Munin, Ganglia, Prometheus*)

Motivation

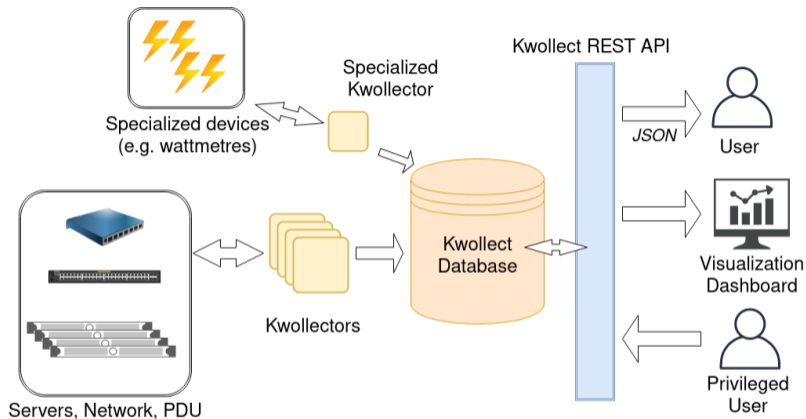
- Monitoring d'infrastructure, focus sur l'environnemental :
 - ▶ Consommation électrique à la prise
 - ▶ Température ambiante
 - ▶ Métrique des BMC, des PDUs
 - ▶ ...
- Grand volume, haute fréquence (*Wattmetre* à 50 mesures/s)
- Conservation longue durée et sans perte possible
- Orienté vers les utilisateurs
 - ▶ Pas un outil pour les admin de la plateforme
 - ▶ API pour récupérer les données, intégration avec le job scheduler.
 - ▶ Métriques "custom"



- Impossible de laisser l'utilisateur se débrouiller seul
- Outils de monitoring traditionnels non adaptés (*Munin, Ganglia, Prometheus*)

→ **Kwollect**

Architecture de Kwollect



PostgreSQL (stockage, API, logique applicative) + Kwollectors + Grafana

→ <https://gitlab.inria.fr/grid5000/kwollect>

Stockage des métriques

- Base de donnée PostgreSQL
 - ▶ + Extension TimescaleDB (auto-partitionnement, compression, agrégation, ...)
 - ▶ + API utilisateur via Postgrest
 - ▶ L'ensemble du «backend» est codé dans la base
- Format des métriques (≈ Prometheus)

Column	Type
timestamp	timestamp with time zone
device_id	text
metric_id	text
value	double precision
labels	jsonb

Dispositifs supportés

Via les *Kwollectors* : programmes autonomes qui récupèrent les métriques et les insèrent dans la base

- Protocoles : SNMP, IPMI, Redfish, Exporteur Prometheus, HTTP / JSON
- Matériel : PDUs, interface d'administration des serveurs, équipements réseau, sondes de température, métrique de l'OS...

+ *Kwollector* spécialisé pour un dispositif donné : *kwollector-wattmetre*

Interface utilisateur

- API : Permet de récupérer les données des métriques
 - ▶ Par date de début/fin, identifiant de noeud/device et de métrique
 - ▶ ... ou par numéro de job.

```
$ curl 'http://kwollect/api/metrics?devices=node-1,node-2&start_time=2020-01-06T13:35:00&end_time=2020-01-06T14:35:00'
```

```
{ "timestamp": "2020-09-11T20:58:49.40357+02:00", "device_id": "node-1", "metric_id": "bmc_gpu_power_watt", "value": 42, "labels": { "gpu": "0" } },  
{ "timestamp": "2020-09-11T20:58:49.40357+02:00", "device_id": "node-1", "metric_id": "bmc_gpu_power_watt", "value": 44, "labels": { "gpu": "1" } },  
{ "timestamp": "2020-09-11T20:58:49.40357+02:00", "device_id": "node-1", "metric_id": "bmc_node_power_watt", "value": 648, "labels": {} },  
{ "timestamp": "2020-09-11T20:58:49.40357+02:00", "device_id": "node-1", "metric_id": "bmc_temp_ambient_celsius", "value": 21, "labels": {} },  
{ "timestamp": "2020-09-11T20:58:50.08682+02:00", "device_id": "node-1", "metric_id": "network_ifaceout_bytes_total", "value": 2654766193, "labels": {} },  
{ "timestamp": "2020-09-11T20:58:50.08682+02:00", "device_id": "node-1", "metric_id": "network_ifacein_bytes_total", "value": 3883940842, "labels": {} }
```

Interface utilisateur

- API : Permet de récupérer les données des métriques
 - ▶ Par date de début/fin, identifiant de noeud/device et de métrique
 - ▶ ... ou par numéro de job.
- Grafana

Interface utilisateur

- API : Permet de récupérer les données des métriques

- Graphe



Quelques chiffres du déploiement Slices-FR/Grid'5000

- Métriques récupérées en continu :
 - ▶ Consommation électrique à la prise, chaque seconde (wattmetre et/ou PDU)
 - ▶ Principales métriques des BMC : température ambiante, consommation à l'alim
 - ▶ Trafic des équipements réseau, chaque seconde
 - ▶ Principales métriques Prometheus (*prometheus-node-exporter* et *dcgm-exporter* pour les GPU Nvidia)
 - ▶ Métriques poussées par les utilisateurs
- À la demande : Mesures de consommation par les wattmetres à 50Hz, toutes les métriques des BMC, toutes les métriques des exporteurs Prometheus, etc.
- Une instance (BDD + Kwolector + ...) déployée sur chacun des 10 sites Grid'5000
 - ▶ Quelques milliers de métriques par seconde sur chacun
 - ▶ Performances : $\approx 1\text{min}$ pour récupérer 10^6 métriques (hébergement sur VMs modestes)

Section 5

Exemple d'utilisation

Exemple d'utilisation dans Slices-FR/Grid'5000

Scénario de l'expérience :

*Etude de la consommation d'un serveur :
au repos et lors de son démarrage + impact de l'exécution d'un service*

- ➊ Réservation d'un nœud dans Grid'5000
 - ▶ Avec activation du monitoring haute fréquence des wattmetres
- ➋ Attente, puis redémarrage du nœud, puis désactivation du service
- ➌ Fin de la réservation
- ➍ Récupération des métriques
- ➎ Analyse

Exécution du scénario

① Réserveation du nœud

```
$ oarsub \  
-p nova \  
-t monitor=wattmetre \  
... # Node HW specification  
... # Enable high freq. on wattmetre
```

② Redémarrage du nœud

```
$ ssh nova-1 "sudo-g5k && sudo reboot"
```

③ Désactivation du service

```
$ ssh nova-1 "sudo systemctl stop ..."
```

④ Fin de la réserveation

Récupération des métriques et analyse

Récupération des métriques via l'API Kwollect

```
$ curl https://api.grid5000.fr/stable/sites/lyon/metrics?job_id=12345 > metrics.json
```

Analyse des résultats en python / pandas

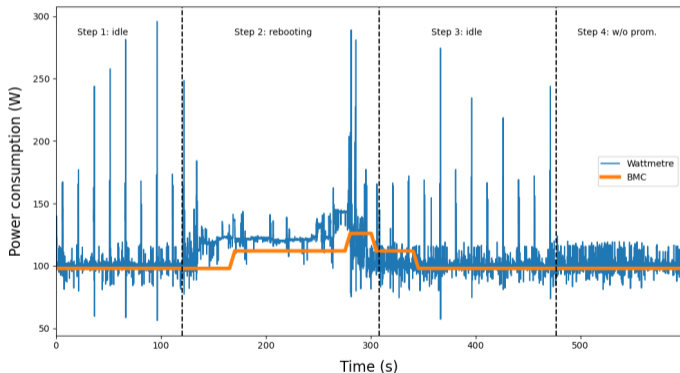
```
import pandas as pd
from matplotlib import pyplot as plt

data = pd.read_json("metrics.json", orient="record")

fig, ax = plt.subplots()
ax.plot(
    data[data["metric_id"] == "wattmetre_power_watt"]["timestamp"],
    data[data["metric_id"] == "wattmetre_power_watt"]["value"],
    label="Wattmetre"
)
ax.plot(
    data[data["metric_id"] == "bmc_node_power_watt"]["timestamp"],
    data[data["metric_id"] == "bmc_node_power_watt"]["value"],
    label="BMC"
)
# ...
plt.savefig("metrics.pdf")
```

Résultats

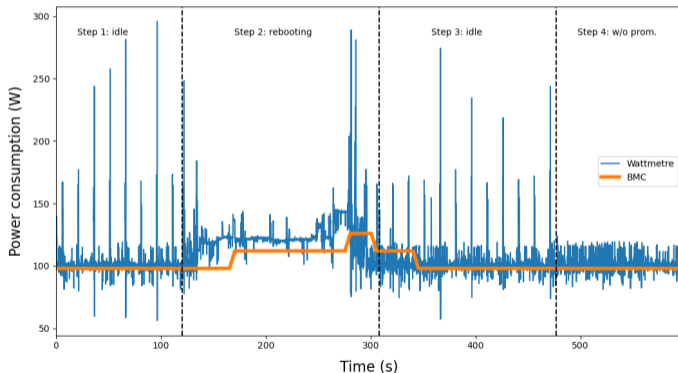
Consommation énergétique d'un nœud, mesuré sur la BMC et avec un wattmètre



- Step 1: Au repos

Résultats

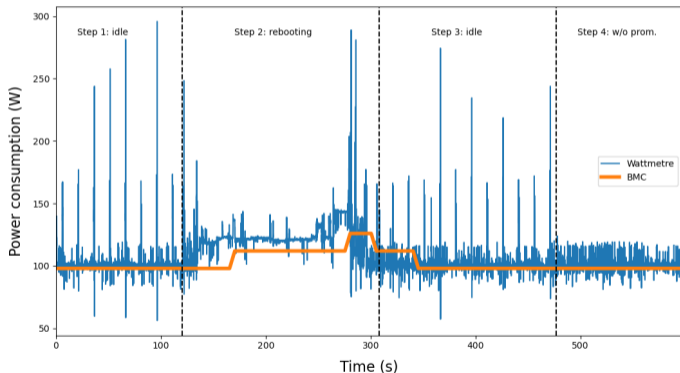
Consommation énergétique d'un nœud, mesuré sur la BMC et avec un wattmètre



- Step 1: Au repos
- Step 2: Reboot de la machine

Résultats

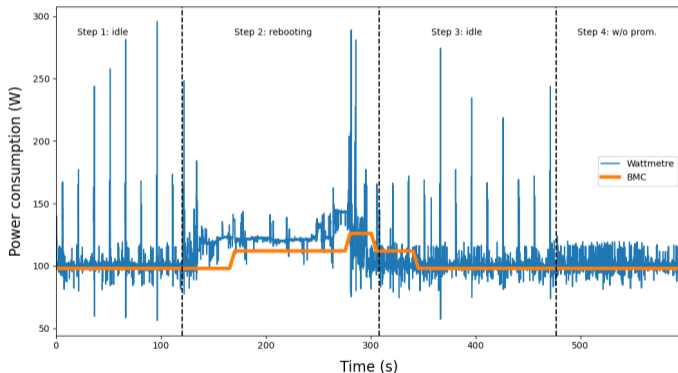
Consommation énergétique d'un nœud, mesuré sur la BMC et avec un wattmètre



- Step 1: Au repos
- Step 2: Reboot de la machine
- Step 3: Au repos

Résultats

Consommation énergétique d'un nœud, mesuré sur la BMC et avec un wattmètre



- Step 1: Au repos
- Step 2: Reboot de la machine
- Step 3: Au repos
- Step 4: Arrêt de l'exporteur Prometheus

Section 6

Conclusion

Conclusion

- Slices-FR/Grid'5000 permet les expérimentations prenant en compte l'énergie
 - ▶ mesures de consommation réelles, à haute fréquence
 - ▶ ⚠ qualité des mesures, confiance dans les résultats
 - ▶ obtenir le coût de son utilisation de la plateforme
- Le framework Kwollect
 - ▶ outil générique pour le monitoring de métriques environnementales et leurs mises à disposition aux utilisateurs des plateformes
 - ▶ aboutissement de nombreuses années d'expérience dans le contexte Grid'5000
- Le futur
 - ▶ Slices : Bientôt possible sur le continuum « du datacenter à l'IoT »
 - ▶ Variété de HW plus importante
 - ▶ Plus de capteurs environnementaux
 - ▶ Au-delà des kW : CO₂, impact environnemental, cycle de vie

Merci !

https://www.grid5000.fr/w/Energy_consumption_monitoring_tutorial

<https://gitlab.inria.fr/grid5000/kwollect>