Offer #2022-05130

PhD Position F/M Energy efficient data management: Data reduction and protection meet performance and energy

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

Context

Financial and working environment.

This PhD will be hosted by Inria (Myriads team, Rennes Bretagne Atlantique) and will be funded by Inria. This sub-project is a part of the Inria-OVH collaborative framework. Thus, the work will be carried out in a close collaboration with OVH. In fact, we plan to validate the results of the project using several OVH data services including backup services and media service, etc.

The PhD student will be supervised by:

- Shadi Ibrahim, member of the Myriads team in Rennes
- Guillaume Pierre, head of the Myriads team in Rennes
- Jean-François Smigielski, Software Engineer specialized in Block Storage, OVHcloud
- Romain De Joux, Technical Lead Object Storage, OVHcloud

Visits and meetings between the successful candidate and the supervisors will be organized, as well as meetings with the other members of the Inria-OVH collaborative framework.

Assignment

Context

The amount of data observed from the world is growing exponentially, reaching 64.2 zettabytes in 2020. To meet the continuously growing demand for computing resources to store and process Big Data, large cloud providers have equipped their infrastructures with millions of energy hungry servers distributed on multiple physically separate data-centers. This results in a tremendous increase in the energy consumed to operate these data-centers. However, as the data and the scale of data-centers are on the rise, energy consumption will continue to be a major concern in the Cloud. Thus, it is important to make data management in the cloud energy-efficient.

Data are usually replicated to ensure high availability and performance (by directing users to the closest replica). However, replication comes with high costs in term storage space, network usage, and performance when writing data. This can be also translated in high energy consumption [1], in particular to store and transfer data.

Recently, we have witnessed advances in the performance of reduction and protection schemes like erasure coding (EC), deduplication, compression, etc. Thus, recent efforts have been dedicated to investigate the potential of replacing replication with erasure coding to reduce the cost of data storage while sustaining good performance. For example, EC is now employed in data analytic systems [2, 3] and in-memory storage systems on cached (hot) data [4]. Though benefits exist, EC poses new challenges including cost of access, energy consumption (encoding, decoding, etc), data availability and data loss. In addition, when adopting EC, we need to take into consideration the frequency and performance requirements of data which vary according to the age and type of data, time of access, the applications, and users.

References

[1] Yacine Taleb, Shadi Ibrahim, Gabriel Antoniu and Toni Cortes: Characterizing performance and energy-efficiency of the ramcloud storage system. In 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), pages 1488–1498, 2017.

[2] Jad Darrous and Shadi Ibrahim: Understanding the performance of erasure codes in hadoop distributed file system. In Proceedings of the Workshop on Challenges and Opportunities of Efficient and Performant Storage Systems (CHEOPS '22). Pages 24–32, 2022.

[3] Jad Darrous, Shadi Ibrahim and Christian Perez: Is it time to revisit erasure coding in data- intensive

clusters ? In 2019 IEEE 27th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), pages 165–178, 2019.

[4] K. V. Rashmi, Mosharaf Chowdhury, Jack Kosaian, Ion Stoica, and Kannan Ramchandran: EC-cache: load-balanced, low-latency cluster caching with online erasure coding. In Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation (OSDI'16).

Main activities

This PhD Thesis will address the problem of how to improve energy efficiency of Big Data services through exploring data reduction and protection schemes (i.e., erasure codes). This research is expected to bring innovative contributions with respect to the following aspects:

- As a first step we need to profile and classify the applications according to their objectives (energy, performance, durability etc.), their access patterns and deployment modes; and *study and model the performance, energy consumption, and data loss of the applications under EC and replication*;
- Data comes with different sizes and has different temperatures (frequency of access), Accordingly, a hybrid scheme (using Replication and EC) is more practical for heterogeneous data (for example, EC may not be the best choice for small files), thus it is essential to evaluate the cost of transforming data between replication and EC when hybrid schemes is used;
- Based on the performance models and the cost model, we will propose innovative data placement and retrieval strategies to optimize the performance and energy consumption of EC that take into consideration the location of users desired performance, the availability of high-speed hardware and the availability of green energy sources.

Skills

- An excellent Master degree in computer science or equivalent
- Strong knowledge of distributed systems
- Knowledge of storage and distributed file systems
- Strong programming skills (C/C++, Python)
- Working experience in the areas of Big Data management, Cloud Computing, Data Analytics are advantageous
- Very good communication skills in oral and written English

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

General Information

- **Theme/Domain** : Distributed Systems and middleware
  System & Networks (BAP E)
- **Town/city** : Rennes
- **Inria Center** : Centre Inria de l'Université de Rennes
- **Starting date** : 2023-10-01
- **Duration of contract** : 3 years
- **Deadline to apply** : 2023-09-30

Contacts

- **Inria Team** : MYRIADS
- **PhD Supervisor** :
  Ibrahim Shadi / Shadi.Ibrahim@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.