Systèmes Multi-niveaux pour l'Optimisation Hybride des Flux dans le Cloud

Stage M2

Contact : Alessio Pagliari, Daniel Wladdimiro, Emmanuel Hyon {prenom.nom@lip6.fr}

Lieu : LIP6, 4 place Jussieu, 75005 Paris

Contexte

Les systèmes de traitement de flux (*Stream Processing Systems* – SPS) dans le Cloud doivent s'adapter à des charges variables et imprévisibles tout en respectant des objectifs de latence (souvent sur des percentiles) et de débit sous contrainte de coûts. Ces applications s'expriment comme des graphes d'opérateurs répartis, où la performance dépend à la fois du parallélisme logique (réplication, partitionnement) et des décisions de déploiement physique (types, quantités et localisation des instances). Les approches classiques d'auto-scaling, fondées sur des heuristiques réactives et un découplage logique/physique, mènent fréquemment soit à une surprovision onéreuse, soit à des congestions et violations de SLA [2, 5].

Sur le plan scientifique, la recherche de configurations efficaces est un problème d'optimisation combinatoire *NP-difficile*, apparenté à des variantes d'ordonnancement (p. ex. *Job Shop*) et rendu plus ardu par la dynamique du système (décisions *en ligne* sous forte contrainte de temps de calcul) [7]. Dans le Cloud, l'hétérogénéité des catalogues d'instances (p. ex. instances *burstable*), les politiques tarifaires (à la seconde, *spot*, egress réseau, stockage) et les contraintes de proximité des sources de données influencent directement le compromis coût–latence–stabilité [6].

Le stage propose un cadre d'optimisation multi-niveaux qui traite conjointement la dimension logique (réplication, partitionnement, placement des opérateurs) et la dimension physique (choix et affectation des instances, topologie Cloud/Edge), avec des décisions en ligne stables et frugales en temps de calcul. L'évaluation s'appuiera sur des plateformes open source (Flink, Storm) et intégrera les coûts Cloud réels dans la fonction objectif (calcul, stockage, trafic sortant), afin de minimiser la dépense tout en respectant des SLA exigeants et en limitant les reconfigurations et migrations coûteuses [3, 1].

Objectifs

L'objectif du projet est de concevoir un modèle d'optimisation conjointe pour les ressources logiques et physiques dans les systèmes de traitement de flux distribués. Les principales étapes incluent :

- Modéliser le problème d'allocation dynamique comme un problème d'optimisation combinatoire [7].
- Intégrer des techniques d'apprentissage par renforcement (*Model-Based RL*) pour anticiper les variations de charge.
- Développer un prototype d'orchestrateur hybride combinant recherche opérationnelle et heuristiques rapides.
- Évaluer expérimentalement les performances sur des plateformes open source (Flink, Storm) déployées sur AWS, GCP ou Azure, en intégrant les coûts réels d'infrastructure [4, 1].

Prérequis

Bonnes connaissances en systèmes distribués et en programmation (Python, Java). Des notions en Cloud Computing, optimisation combinatoire ou apprentissage automatique sont un atout. Une expérience préalable avec Flink, Storm ou Kubernetes est souhaitable.

Références

[1] Farah Aït-Salaht, Laëtitia Della Maestra, and Daniel Wladdimiro. ACADS : A Framework for Adaptive Cost-Aware Deployment of Stream Processing System in the Cloud. In *13th IEEE International Conference on Cloud Engineering*, Rennes, France, September 2025.

[2] Valeria Cardellini, Francesco Lo Presti, Matteo Nardelli, and Gabriele Russo Russo. Runtime adaptation of data stream processing systems : The state of the art. *ACM Comput. Surv.*, 54(11s) :237 :1–237 :36, 2022.

[3] Alessio Pagliari and Guillaume Pierre. Transscale : Combined-approach elasticity for stream processing in fog environments. pages 17–24, 2023.

[4] A. Poobalan, P. Shanthakumar, and M. Robinson Joel. Hybrid optimization enabled VM scaling based load distribution and optimal switching strategy in cloud data center. *Wirel. Networks*, 30(2) :1085–1105, 2024.

[5] Samuel Rac and Mats Brorsson. Cost-effective scheduling for kubernetes in the edge-to-cloud continuum. In *IEEE International Conference on Cloud Engineering, IC2E 2023, Boston, MA, USA, September 25-29, 2023*, pages 153–160. IEEE, 2023.

[6] Daniel Wladdimiro, Alessio Pagliari, and Rafaela C Brum. Toward Stream Processing Efficiency Leveraging Cloud Burstable Instances. In *13th IEEE International Conference on Cloud Engineering*, Rennes, France, September 2025.

[7] Hegen Xiong, Shuangyuan Shi, Danni Ren, and Jinjin Hu. A survey of job shop scheduling problem : The types and models. *Computers & Operations Research*, 142 :105731, 2022.