Proposition de thèse PEPR eNSEMBLE

Collaboration humain-agents intelligents : enjeux environnementaux des infrastructures contraintes

Encadrants :
- Thomas Ledoux (HDR), IMT Atlantique
- Dimitri Saingre, IMT Atlantique

Laboratoire d'accueil : LS2N, site de IMT Atlantique à Nantes

Court résumé : Cette thèse s'intéresse à l'impact environnemental croissant des agents intelligents basés sur les grands modèles de langages, dont le développement entraîne une escalade jugée insoutenable des besoins en infrastructure numérique. Elle propose d'évaluer les apports d'une approche alternative exploitant uniquement du matériel existant (ordinateurs sans carte graphiques, cartes graphique "grand public" ou d'anciennes générations, …) dans un cadre d'infrastructure numérique partagée. La thèse vise d'une part à cartographier les problématiques techniques freinant l'utilisation de matériel informatique d'anciennes générations et d'autre part à proposer des pistes de solutions visant à permettre l'utilisation d'agents intelligents sur des infrastructures hétérogènes et d'ancienne génération. Ces pistes viseront à définir un cadre de collaboration entre , d'une part, l'utilisateur et ses besoins, et, d'autre part, et la plateforme d'hébergement et son objectif d'auto-limitation en ressources. Enfin, elle vise à caractériser l'impact environnemental net d'un tel scénario où le déploiement de ces agents serait conditionné majoritairement à des infrastructures numériques basées sur du matériel grand public.

Description du groupe de recherche / laboratoire d'accueil : Cette thèse sera encadrée par l'équipe STACK (IMT Atlantique, Inria, Nantes Université). L'équipe STACK est spécialisée dans la gestion d'infrastructures de services centralisées (Cloud) et décentralisées (continnum entre l'Edge et le Cloud). L'équipe aborde des thèmes variés comme la gestion des ressources (matériel et énergie), les middleware, les langages et applications, etc. A travers des travaux sur la mesure et la maîtrise du coût énergétique de services numériques et la recherche de compromis entre performance et empreinte, l'équipe STACK cherche aussi à proposer des infrastructures plus sobres.

1 Description de la proposition de doctorat

1.1 Contexte et scénario

Le récent développement des technologies d'*intelligence artificielle génératives* dans divers domaines d'application pose des problèmes environnementaux importants. Ces technologies prennent la forme d'*agents intelligents*, construits à partir de grands modèles de langage (LLM, pour la génération de texte) ou de modèles de diffusion (pour la génération d'image ou vidéo).

Ces technologies (ci-après nommées *agents intelligents*) nécessitent bien souvent l'utilisation d'infrastructures coûteuses, que ce soit lors de leur phase d'entrainement ou d'inférence [1].

Nous observons donc aujourd'hui une réelle escalade des besoins en infrastructure numérique (des composants aux centres de données) liée à au développement de ces LLMs. Cette escalade est considérée ici comme indésirable, voire insoutenable sur plusieurs aspects :

- **Environnemental :** la construction de nouveaux serveurs et la construction de nouveaux centres de données pour les héberger présente un impact environnemental majeur
- **Economique :** la demande croissante en composants informatique induit une hausse significative des prix, aujourd'hui constatée par exemple sur la mémoire RAM
- **Souveraineté :** l'utilisation de grand modèles de langage nécessite aujourd'hui souvent des investissements majeurs en infrastructure numérique. Nous voyons cela comme un facteur de dépendance des utilisateurs à de grands industriels disposants de cartes graphiques coûteuses, limitant l'hébergement de ces outils par de plus petits acteurs (petites entreprises, auto-hébergement, infrastructures communautaires…)

1.2 Problématique

Le point de départ de cette thèse est l'exploration d'un scénario où utilisation d'agents intelligents se conjuge avec le refus de la construction de nouvelles infrastructures numérique. Dans ce cadre, nous considérons le déploiement de services d'inférence sur des infrastructures :

- **Hétérogènes** : serveurs aux performances hétérogènes et composés de processeurs et accélérateurs (ou non) de différentes architectures et différentes générations
- **Contraintes** : le non-renouvellement des infrastructures impose une limite naturelle en terme de capacité correspondant à l'existant
- **Collaboratives** : nous considérons des scénarios de partage de ressources dans un cadre visant à inclure les utilisateurs et leurs besoins dans la réduction de l'empreinte environnementale de leurs agents intelligents.

De telles infrastructures ne sont aujourd'hui que peu considérées pour l'hébergement d'agents intelligents. La principale raison est le besoin important de ressources de calculs nécessaire par ce type d'application. De fait, ce scénario s'inscrit en quelques sorte en opposition de ce que nous observons actuellement, où ces modèles de langages sont majoritairement déployés sur de larges infrastructures homogènes et possédées par un nombre d'acteurs industriels restreint.

1.3 Bref état de l'art et fondements théoriques

L'empreinte environnemental du numérique est en constante augmentation et représente déjà aujourd'hui un impact significatif [2]. Traditionnellement, la majeure partie de cette empreinte est liée à la construction de matériel informatique [2] et à l'extraction de ressources première associée [3]. Le développement des technologies d'IA générative contribue à augmenter cette empreinte, d'une part pour la fabrication de nouvelles infrastructure et d'autre part à cause du coût énergétique important nécessaire pour le fonctionnent (appelé *inférence*) des modèles d'IA [4].

En réponse, de nombreux travaux ont étudié les leviers matériels et logiciels pouvant réduire le coût en ressources de ces technologies. A titre d'exemple, sur le plan matériel, le pilotage de la fréquence d'exécution des accélérateurs permet de réduire leur consommation d'énergie au prix d'une latence plus élevée [5]. Sur le plan logiciel, la sélection d'un modèle plus petit, la quantisation ou les options de parallélisation permettent de réduire le coût en ressources de ces modèles [6], [7]. De ces études, des solutions améliorant le pilotage de services d'inférences ont émergé[5], [8]. Ces solutions visent souvent à reconfigurer dynamiquement la plateforme pour obtenir le meilleur ratio coût - performances. Cependant, ces travaux se basent exclusivement sur des infrastructures modernes, coûteuses et à très hautes performances. Nous n'avons trouvé que très peu de travaux sur les contraintes liées à l'utilisation de matériel informatique grand public.

Le domaine de recherche *computing within limits* [9] s'intéresse à l'impact des limites planétaires sur l'informatique. Un levier d'action possible est l'utilisation de matériel de seconde main. Certains travaux de recherche explorent des scénarios poussé avec le ré-emploi de *smartphones* usagés pour l'hébergement d'application web [10]. Ces travaux peuvent servir de source d'inspiration au regard des possibilités liées au réemploi de matériel usagé. Pour permettre de limiter les besoins en ressources de services numériques, une approche consiste à identifier des leviers fonctionnels, actionable par les utilisateurs de service numérique. Madon et al. [11] identifient quatre leviers fonctionnels : le délai, la dégradation des résultats, la reconfiguration et le renoncement. D'autres travaux s'intéressent à l'implication des utilisateurs dans la réduction d'empreinte environnementale, que ce soit via la présentation de *feedback* [12] ou la présentation de leviers d'actions [13]. Des leviers de qualité de service [14], [15] ont été explorés par les encadrants de ce sujet de thèse et montrent que des leviers relatifs à la qualité de l'expérience utilisateur peuvent conduire à une diminution significative du coût de fonctionnement d'un service numérique. Impliquer des utilisateurs à la réduction de l'empreinte des services utilisés apparait comme une opportunité intéressante pour aller au delà de l'optimisation technique.

Du côté des technologies d'IA générative, Quelques initiatives telles que *CanIRun.ai*[1] visent à collecter des données permettant d'estimer les performances d'un modèle sur un matériel donné, même si les facteurs expliquant ces performances sont encore peu étudiés. Le projet FrugalIA[2] auquels participent les encadrants de ce projet de thèse vise à mesurer le coût environnemental d'agents intelligents et à le limiter en recommandant les agents les moins coûteux pouvant satisfaire un besoin donné. Cependant, la définition de mécanismes et de protocoles de collaborations visant spécifiquement l'auto-limitation en ressources reste un terrain de recherche sous-exploité.

1.4 Questions de recherche

Ce projet de thèse vise à étudier la faisabilité et l'impact environnemental de l'utilisation d'agents d'IA générative dans un contexte encore sous étudié : l'utilisation de matériel contraint et grand public dans un cadre de collaboration avec les usagers. En particulier, nous proposons d'adresser les questions de recherche suivantes, encore aujourd'hui sous-étudiées :

- Quelles sont les barrières techniques limitant l'utilisation d'agents intelligents sur une infrastructure *d'ancienne génération* ? Comment caractériser la compatibilité d'un modèle d'intelligence artificielle générative sur une infrastructure donnée ?
- Quels modèles de collaboration permettent de rendre un service d'agent intelligent souhaitables aux utilisateurs tout en limitant son impact à un cadre soutenable ?
- Sous quelles conditions, l'utilisation d'infrastructure contraintes et l'auto-limitation permettent-elles de rendre l'utilisation d'agents intelligents soutenables d'un point de vue environnemental ?

[1] https://www.canirun.ai/
[2] https://frugalia.eu/

1.5 Approche, objectifs et évaluations

Ce projet de thèse adoptera principalement une approche empirique et expérimentale, à travers trois grands objectifs.

Le premier objectif de cette thèse sera d'établir un diagnostique sur d'une part les facteurs définissants la compatibilité d'un modèle vis à vis d'un matériel (capacités en ressources, architecture, bande passante, ...) et d'autre part les leviers directement actionnable pour améliorer la compatibilité d'un modèle vis à vis d'un ensemble de machines. A titre d'exemple, la quantisation est un mécanisme permettant de réduire les exigences en mémoire d'un modèle. Ces premiers travaux sont nécessaires pour dresser les limites de ce qui sera envisageable par la suite.

Cette étude s'effectuera sur une infrastructure réelle, dans un premier temps sur Grid5000[3]. Grid5000 offre un panel de serveur aux spécifications très variées. Il sera alors possible d'évaluer des modèles de langages sur des serveurs de plus en plus anciens et contraints. Nous envisageons dans un second temps la collecte d'anciens ordinateurs portable pour inclure des terminaux utilisateurs dans ces travaux.

Partant de ce diagnostique, le second objectif de cette thèse sera l'évaluation de leviers techniques et fonctionnels permettant d'améliorer la compatibilité de modèles d'IA pour une infrastructure contrainte et de limiter la taille de cette infrastructure via des mécanismes de partage de ressources. Ces différents leviers seront évalués sur des axes d'impacts en terme de performance (ex : temps de réponse), de consommation de ressources (ex : énergie et nombre de serveurs) et de pertinence des résultats proposés par les modèles. Cette étude pourra s'effectuer sur infrastructure réelle, en prenant pour base des infrastructures décentralisés tels que [16], ou sur un simulateur tel que SimGrid [17]. Ce projet vise à prioriser des leviers permettant à aux utilisateurs d'agents intelligents de participer à la maîtrise du coût en ressources de leur service.

Enfin, le dernier objectif de cette thèse sera l'évaluation de l'impact environnemental de l'utilisation d'agents intelligents dans un cadre d'infrastructure contrainte. A partir du diagnostique réalisé lors de la première phase et des leviers évalués lors de la seconde phase, des scénarios comparatifs seront définis. L'impact environnemental de ces différents scénario seront évalués en se basant sur les méthodologies aujourd'hui utilisés pour étudier l'impact environnemental d'outils d'intelligence artificielle, souvent basés sur des méthodes type *analyse du cycle de vie*.

En s'inscrivant en opposition aux développement et déploiement continu de nouvelles infrastructures et de nouvelles générations de serveurs, ces travaux de thèse viseront à challenger cette escalade numérique et à rendre accessible les outils d'IA génératives sur du matériel plus ancien.

Pour limiter le sujet de cette thèse à un cadre réalisable, nous mettons de côté les sujets liés à l'entrainement de ces agents et à une réadaptation *profonde* des modèles d'IA (nouvelles architectures, nouveaux moteurs d'inférence, ...). Cependant, les travaux conduits dans cette thèse pourront servir de point de départ pour la construction de modèles d'IA utilisés **et** entraînés sur des infrastructures contraintes.

[3] https://www.grid5000.fr/

2 Nature de la collaboration numérique

Dans le cadre de ce projet de thèse, nous cherchons à évaluer deux axes de collaboration entre des utilisateurs et une plateforme délivrant un service d'agent intelligent.

Le premier axe de collaboration envisagé est entre un **usager** et la **plateforme** hébergeant les agents intelligents. D'un point de vue environnemental, cette collaboration nous semble cruciale pour pouvoir dépasser les possibilités offertes par la simple optimisation logicielle. Pour la plateforme, l'objectif de cette collaboration et de pouvoir délivrer un service d'agent intelligent en limitant le plus possible son coût environnemental (de la mobilisation de matériel informatique à sa consommation de ressources lors du fonctionnement). Pour l'utilisateur, l'objectif est de pouvoir bénéficier d'un service d'agent intelligent avec des contraintes de soutenabilité environnementales fortes. Une coordination sera nécessaire entre les deux acteurs pour arriver à un point de compromis entre d'une part les besoins de l'utilisateur et d'autre part les objectifs d'auto-limitation de la plateforme. Cette collaboration peut-être rendue possible en rendant les utilisateurs acteurs de leur consommation de ressources. Il est possible pour cela de présenter aux utilisateurs des retours sur l'état de fonctionnement de la plateforme ainsi que différents leviers (identifiés et évalués lors du projet) pouvant influer sur la consommation en ressources de leurs agents intelligents. Le cadre d'usage (l'utilisation d'agents intelligents) dans lequel s'inscrit ce sujet de thèse, permet la mise en place de leviers fonctionnels variés (ex: génération asynchrone de contenu au lieu d'une réponse instantannée, dégradation de la qualité de réponse, ...). En choisissant ou non d'actionner ces leviers l'utilisateur engage alors une relation de collaboration avec la plateforme en construisant des compromis entre ses besoins en terme d'usages et ses besoins en termes de ressources.

Le second axe de collaboration se place **entre** les utilisateurs d'une même plateforme. Nous considérons ici une collaboration indirecte, où une communauté d'utilisateurs (entreprise, association, ...) collabore via la plateforme pour atteindre des besoins globaux en minimisant la consommation de ressources (but de la plateforme ici aussi). Des plateformes partagées comme Grid5000[4] rendent explicite le partage de ressources en fixant des règles d'utilisation et rendant visible la disponibilité de chaque serveur réservable. Il nous semble possible d'aller plus loin. Les travaux de la littérature visant à inclure l'utilisateur dans sa réduction de consommation de ressources informatique s'inscrivent majoritairement dans une collaboration unique entre un utilisateur et un service numérique. Nous souhaitons ici étendre ces travaux pour permettre à une communauté d'utilisateurs de réduire les besoins en ressources d'un service numérique partagé sans nécessiter une collaboration explicite / directe entre les utilisateurs.

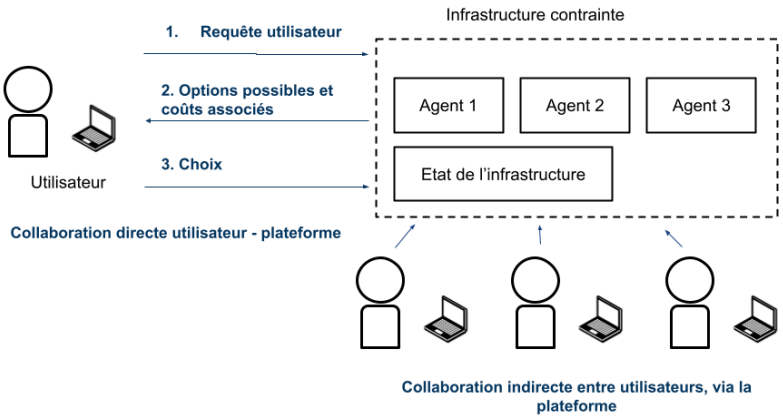

Figure 1: Deux axes de collaboration considérés

[4]https://www.grid5000.fr

3 Contribution à la collaboration numérique : Résultats attendus et impact

Ce projet de thèse vise à adresser un défi d'actualité qui est celui de l'impact environnemental majeur lié au développement d'outils d'agents basés sur des IA génératives. En se positionnant dans un cadre restreint d'infrastructure contrainte, il vise à rendre ces outils accessibles à l'auto-hébergement et soutenables. Evoluer dans ce cadre encore sous-étudié pourra mener à plusieurs contributions d'intérêts pour la communauté.

Nous identifions un axe de contribution empirique lié à l'identification des freins limitant l'utilisation de matériel informatique d'ancienne génération. Comprendre ces mécanismes est crucial pour pouvoir améliorer la mise en oeuvre d'agents intelligents sur du matériel accessible et grand publique.

Un second axe de contribution, méthodologique, consiste en la construction d'un protocole de collaboration entre usagers et plateforme d'hébergement pour permettre l'utilisation d'un service dans un cadre d'infrastructure contrainte. Nous anticipons que des leviers techniques (optimisation, reconfiguration, ...) ne sauraient suffire. Développer des leviers fonctionnels (permettre par exemple aux utilisateurs de spécifier un délai d'attente jugé raisonnable pour limiter des effets de pics de charge) apparaît nécessaire dans le cadre que nous nous imposons. Partant de là, nous souhaitons comprendre comment passer de ces différents leviers à une collaboration efficace entre les utilisateurs et leurs besoins d'une part et la plateforme et ses objectifs d'auto-limitation d'autre part. A terme, l'objectif est de repenser la construction de services numériques en placant l'auto-limitation en ressources au coeur de ces services (en rendant la consommation de ressource informatique *tangible* et *actionnable*). Cet axe de contribution méthodologique pourra s'étendre en une contribution technique via le développement d'un prototype pouvant être exploité dans de futurs projets, par exemple dans le cadre de tests utilisateurs.

Enfin, un dernier axe de contribution théorique consistera en une prise de recul sur l'impact environnemental de ce type d'approche basée sur du matériel contraint. Cette contribution pourra prendre en comparaison le coût environnemental d'une plateforme basée sur une infrastructure contrainte et celui d'une infrastructure "classique", représentative de celles déployées aujourd'hui. Au regard de la question environnementale, nous anticipons que ce type d'approche pourra présenter des opportunités (limitation de la construction de nouvelles infrastructures, auto-limitation vis à vis des cas d'usage les plus consommateurs) mais aussi des limites (utilisation de matériels peut-être moins efficients). Si les deux premiers axes visent à étudier des problématiques de faisabilité techniques (freins à l'utilisation de matériel grand public) et fonctionnels (leviers de collaborations entre humain et service dans un cadre de ressources contraintes), ce dernier axe vise à permettre la construction d'une image complète de l'impact environnemental de l'utilisation d'infrastructures contraintes dans le cadre d'une collaboration entre humain et agents intelligents.

Au regard de la question environnementale, l'objectif de ce projet de thèse n'est pas tant la mesure du coût environnementale de l'utilisation d'agent intelligent que la redéfinition de services numériques (aujourd'hui à forte empreinte environnementale), dans un cadre d'auto-limitation en ressources via des protocoles de collaboration entre les utilisateurs et la plateforme hébergeants ces services.

4 Positionnement dans le programme eNSEMBLE

Ce projet de thèse s'intéresse au coût environnemental d'une collaboration à deux échelles : d'une part entre utilisateurs et agents intelligents et d'autre part entre les utilisateurs et les infrastructures hébergeants ces agents.

Il propose de conduire des études sur le coût de l'utilisation de ces systèmes intelligents dans un cadre particulier d'infrastructure contrainte. Ce cadre présente deux intérêts. Le premier est de rendre accessible l'auto-hébergement de systèmes intelligents à différentes communautés. Cela est nécessaire pour des enjeux de souveraineté et pour permettre à toutes communautés de s'approprier et d'avoir la réelle maîtrise de ces outils. Le second est la limitation de l'impact environnemental de ces systèmes en freinant la fabrication de nouvelles infrastructures.

En considérant des contraintes environnementales fortes et en mesurant l'impact liés à l'utilisation de systèmes intelligents, ce projet correspond aux objectifs scientifiques du PC5 TRANSVERSE. En particulier, il raisonne avec ses thèmes 1 et 3.

Ce projet raisonne aussi avec les travaux conduit dans le PC3 MATCHING en s'intéressant au coût de systèmes numériques impliquant une relation entre humain et agents intelligents. Les travaux portés dans le cadre du PC3 pourront potentiellement être intégrés comme cas d'usage dans les études conduites par ce projet. A l'inverse, ce projet pourra venir en soutien pour la mise en place d'expérimentations et la mesure du coût environnemental des projets du PC3.

Bibliographie

[1] A. Berthelot, E. Caron, M. Jay, and L. Lefèvre, "Estimating the environmental impact of Generative-AI services using an LCA-based methodology," *Procedia CIRP*, vol. 122, pp. 707–712, 2024.

[2] Y. Aiouch, A. Chanoine, L. Corbet, P. Drapeau, L. Ollion, and V. Vigneron, "Evaluation de l'impact environnemental du numérique en France et analyse prospective, Etat des lieux et pistes d'actions.," 2022.

[3] S. Cerf, A. Luxey-Bitri, C. Quinton, R. Rouvoy, T. Simon, and C. Truffert, "Untangling the Critical Minerals Knot: when ICT hits the Energy Transitions," 2023.

[4] IEA, "Energy and AI," 2025.

[5] J. Stojkovic *et al.*, "Tapas: Thermal-and power-aware scheduling for LLM inference in cloud platforms," in *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 2025, pp. 1266–1281.

[6] J. Fernandez, C. Na, V. Tiwari, Y. Bisk, S. Luccioni, and E. Strubell, "Energy considerations of large language model inference and efficiency optimizations," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 32556–32569.

[7] C. Niu, W. Zhang, Y. Zhao, and Y. Chen, "Energy efficient or exhaustive? benchmarking power consumption of llm inference engines," *ACM SIGENERGY Energy Informatics Review*, vol. 5, no. 2, pp. 56–62, 2025.

[8] J. Stojkovic, C. Zhang, Í. Goiri, J. Torrellas, and E. Choukse, "Dynamollm: Designing llm inference clusters for performance and energy efficiency," in *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2025, pp. 1348–1362.

[9] B. Nardi *et al.*, "Computing within limits," *Communications of the ACM*, vol. 61, no. 10, pp. 86–93, 2018.

[10] J. Switzer, G. Marcano, R. Kastner, and P. Pannuto, "Junkyard computing: Repurposing discarded smartphones to minimize carbon," in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 2023, pp. 400–412.

[11] M. Madon, G. Da Costa, and J.-M. Pierson, "Characterization of different user behaviors for demand response in data centers," in *European Conference on Parallel Processing*, 2022, pp. 53–68.

[12] T. Zaragoza, T. Soulance, A. Noureddine, and E. Exposito, "Understanding and Influencing End-User Behavior in Software Energy Consumption," in *Proceedings of the 29th International Conference on Evaluation and Assessment in Software Engineering*, 2025, pp. 546–556.

[13] D. Guyon, A.-C. Orgerie, C. Morin, and D. Agarwal, "Involving users in energy conservation: A case study in scientific clouds," *International Journal of Grid and Utility Computing*, vol. 10, no. 3, pp. 272–282, 2019.

[14] A. Mokhtari, B. Jonglez, and T. Ledoux, "Towards digital sustainability: involving cloud users as key players," in *2024 IEEE International Conference on Cloud Engineering (IC2E)*, 2024, pp. 126–132.

[15] L. Gazeau and T. Ledoux, "Feature Degradation for Frugality: A Case Study of Overleaf Application," in *GreenArch 2026-1st ICSA workshop on Software Architecture for Green Sustainable Carbon-aware Software Systems*, 2026.

[16] A. Alidra *et al.*, "SeMaFoR - Self-Management of Fog Resources with Collaborative Decentralized Controllers," in *2023 IEEE/ACM 18th Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*, 2023, pp. 25–31. doi: 10.1109/SEAMS59076.2023.00014.

[17] H. Casanova, "Simgrid: A toolkit for the simulation of application scheduling," in *Proceedings First IEEE/ACM International Symposium on Cluster Computing and the Grid*, 2001, pp. 430–437.