



**MINISTÈRE
DE L'ÉDUCATION
NATIONALE,
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE**

*Liberté
Égalité
Fraternité*

Secrétariat général

Direction du numérique
pour l'éducation
Sous-direction des services
numériques
Bureau des services et outils
numériques pour l'éducation
(DNE SN1)

99, rue de Grenelle
75357 Paris SP 07

Secrétariat général
Service de l'action
administrative et des
moyens
Sous-direction des achats
(SAAM B)
Bureau de la stratégie
et de l'ingénierie des achats
(SAAM B1)

61-65, rue Dutot
75732 Paris Cedex 15

CAHIER DES CLAUSES TECHNIQUES PARTICULIÈRES

ANNEXE 05.2 : Rapport des tests de Charge MYSQL

Procédure : MEN-SG-AOO-24002

Objet : Prestations de prise en charge de la solution du gestionnaire d'accès aux ressources (GAR), d'hébergement, d'exploitation, de maintenance, de support et de développement de ladite solution pour le compte du ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche.

RENATER - GAR - Bench

Rapport tests de performance R7.1.90

Le 20/10/2023

Évolutions successives

Version	Date	Description	Auteur(s)
1.0	20/10/2023	Création	Worldline

Table des matières

1	Introduction	5
1.1	Objet du document	5
1.2	Responsabilités liées au document	5
1.3	Documents de référence.....	5
1.4	Autres documents utilisés	5
1.5	Abréviations.....	5
1.6	Glossaire.....	6
1.7	Définition du cadre pour ce test en charge	7
2	Résumé	8
2.1	Stratégie de bench	8
2.2	Déroulement des tests de performances.....	9
2.2.1	Equipes Worldline.....	9
2.2.2	Applications	10
2.2.3	Configuration plateforme et volumétrie	Erreur ! Signet non défini.
2.3	Bilan des résultats des tirs multi instances.....	11
3	Présentation des résultats des tirs de performance	14
3.1	IHM Affectation	14
3.1.1	Description du tir	14
3.1.2	Stratégie des tirs	14
3.1.3	Optimisations.....	14
3.1.4	Tirs mono instance.....	15
3.1.1	Conclusion tirs mono instance.....	17
3.1.2	Tirs multi instances	17
3.1.3	Conclusion tirs multi instances.....	20
3.1.4	Focus sur la volumétrie de production de l'année 2019/2020.....	20
3.2	WS Liste ressources	21
3.2.1	Description du tir	21
3.2.2	Stratégie des tirs	21
3.2.1	Optimisations.....	21
3.2.2	Tirs multi instances	21
3.2.3	Conclusion tirs multi instances.....	23
3.3	Accès ressources.....	24
3.3.1	Description du tir	24
	3.3.2	Stratégie

du tir.....	24
3.3.3 Optimisations.....	24
3.3.4 Tirs mono instance.....	24
3.3.5 Conclusion tirs mono instance.....	27
3.3.6 Tirs multi instances.....	28
3.3.7 Conclusion tirs multi instances.....	30
3.4 Collecte et import des données ENT.....	30
3.4.1 Description du tir.....	30
3.4.2 Stratégie du tir.....	31
3.4.3 Optimisations.....	31
3.4.4 Tirs import FULL en mono instance.....	31
3.4.5 Conclusion import FULL mono instance.....	33
3.4.1 Tirs import DELTA en mono instance.....	33
3.4.2 Conclusion import DELTA mono instance.....	35
3.4.1 Tirs import FULL en multi instance.....	35
3.4.2 Conclusion import FULL multi instance.....	37
3.4.3 Tirs import DELTA en multiinstance.....	37
3.4.4 Conclusion import DELTA multi instance.....	39
3.5 Pré-affectation établissement.....	39
3.5.1 Description du tir.....	39
3.5.2 Stratégie du tir.....	39
3.5.3 Tir mono instance.....	39
3.5.4 Conclusion du tir mono instance.....	41
3.6 Pré-affectation niveau éducatif.....	41
3.6.1 Description du tir.....	41
3.6.2 Stratégie du tir.....	41
3.6.3 Tirs mono instance.....	41
3.6.4 Conclusion du tir mono instance.....	43
3.7 Affectation nouvel arrivant dans un établissement.....	43
3.7.1 Description du tir.....	43
3.7.2 Stratégie du tir.....	43
3.7.3 Tirs mono instance.....	44
3.7.4 Conclusion du tir mono instance.....	45

1 Introduction

1.1 Objet du document

Ce document a pour objectif de présenter la synthèse générale ainsi que les résultats des tests en charge des composants du GAR.

1.2 Responsabilités liées au document

Worldline est responsable de la rédaction de ce document.

1.3 Documents de référence

Numéro	Réf. Document	Type

1.4 Autres documents utilisés

Numéro	Version et/ou Date	Réf. Document	Type
1			
2			

1.5 Abréviations

Abréviation	Signification

1.6 Glossaire

La suite du document s'appuie sur les termes issus du glossaire CFTL disponible publiquement à l'adresse suivante :

<http://www.cftl.fr/wp-content/uploads/2015/03/Glossaire-des-tests-de-logiciel-2-2-F-P1.pdf>

Ci-dessous un extrait des termes et définitions utilisés dans la suite du document

Terme	Signification
Test de performance	Le processus de test pour déterminer les performances d'un produit logiciel.
Fuite mémoire	Une défaillance d'accès à la mémoire causée par un défaut dans la logique d'allocation dynamique de l'espace de stockage d'un programme. Cette défaillance fait que le programme ne libère pas la mémoire quand il a fini de l'utiliser, causant au bout du compte la défaillance de ce programme et/ou d'autres processus concurrents par manque de mémoire.
Bouchon	Une implémentation spéciale ou squelettique d'un composant logiciel, utilisé pour développer ou tester un composant qui l'appelle ou en est dépendant. Cela remplace un composant appelé. [d'après IEEE 610]
Environnement de test	Environnement contenant le matériel, les instruments, les simulateurs, les outils logiciels et les autres éléments de support nécessaires à l'exécution d'un test [d'après IEEE 610]
Objectif de test	Une raison ou but de la conception et l'exécution d'un test.
Test d'endurance	Un type de test mené pour évaluer, sur une période relativement longue, le comportement d'un composant ou système avec une charge constante. Objectif : s'assurer du bon fonctionnement de l'application en condition d'utilisation réelle et sur la durée (p.ex pour vérifier l'absence de fuites mémoire)
Test de charge	Un type de test mené pour évaluer le comportement d'un composant ou système avec une charge croissante, p.ex. nombre d'utilisateurs et/ou nombre de transactions en parallèle pour déterminer quelle charge maximale peut être gérée par le composant ou système. Objectif : identifier le (ou les) point(s) de contention vis-à-vis des hypothèses de temps de réponse
Test de stress	Un type de test mené pour évaluer le comportement d'un composant ou système au-delà des limites de ses charges de travail anticipées ou spécifiées, ou avec une disponibilité réduite de ressources telles que l'accès mémoire ou serveur [d'après IEEE 610]. Objectif : identifier la capacité maximale de l'infrastructure technique (p.ex CPU à 100%)
Neoload	Outils de tests de performance (http://www.neotys.fr/neoload/overview)
Hit	c'est un accès à un unique élément au sens http du terme. Dans le cas des accès Web un hit ne représente qu'une composante de la page. L'accès à une page des accès Web est donc la composition des hits de l'ensemble des éléments qui la compose. Dans le bilan, nous utiliserons davantage le terme de « débit » plutôt que le nombre de Hit/s. Lorsqu'il le sera possible, nous parlerons d'accès/s ou de demande/s pour simplifier la lecture des analyses.
Temps de réponse	le temps de réponse d'une requête ou d'une page est le temps qui s'est écoulé entre l'émission de la requête et la réception de la réponse.
Percentile / Centile	En statistique descriptive, un percentile est chacune des 99 valeurs qui divisent les données triées en 100 parts égales, de sorte que chaque partie représente 1/100 de l'échantillon de population.

1.7 Définition du cadre pour ce test en charge

Ce test en charge a été exécuté dans le cadre de la release 7.1.90 du GAR afin d'analyser l'impact de la refonte du batch et de la montée de version MySQL (5.6 ->8) du schema GAR-ENT

La liste des services identifiés pour ce test ainsi que le(s) type(s) de tir sont définis dans le tableau suivant :

Services	Types de tir	A exécuter
Collecte des données ENT	Performance/Stabilité/Endurance (Mono service Mono instance)	<input checked="" type="checkbox"/>
	Performance/Stabilité/Endurance (Mono service Multi instances)	<input checked="" type="checkbox"/>
Import des données ENT	Performance/Stabilité/Endurance (Mono service Mono instance)	<input checked="" type="checkbox"/>
	Performance/Stabilité/Endurance (Mono service Multi instances)	<input checked="" type="checkbox"/>
Batch de pré-affectations	Performance/Stabilité/Endurance (Mono service Mono instance)	<input checked="" type="checkbox"/>
	Performance/Stabilité/Endurance (Mono service Multi instances)	<input checked="" type="checkbox"/>
WS Liste ressources	Montée en charge (Mono service Mono instance)	<input checked="" type="checkbox"/>
	Performance/Stabilité/Endurance (Mono service Mono instance)	<input checked="" type="checkbox"/>
	Performance/Stabilité/Endurance (Mono service Multi instances)	<input checked="" type="checkbox"/>
Accès ressources	Montée en charge (Mono service Mono instance)	<input checked="" type="checkbox"/>
	Performance/Stabilité/Endurance (Mono service Mono instance)	<input checked="" type="checkbox"/>
	Performance/Stabilité/Endurance (Mono service Multi instances)	<input checked="" type="checkbox"/>
IHM Affectations	Montée en charge (Mono service Mono instance)	<input checked="" type="checkbox"/>
	Performance/Stabilité/Endurance (Mono service Mono instance)	<input checked="" type="checkbox"/>
	Performance/Stabilité/Endurance (Mono service Multi instances)	<input checked="" type="checkbox"/>

2 Résumé

2.1 Stratégie de bench

La stratégie de bench est détaillée dans le document **Erreur ! Source du renvoi introuvable.**

En résumé, la stratégie pour ces tests de performance est la suivante :



- 1 Bench de mesure de la limite de performance pour chaque service
 - Objectif : déterminer le maximum d'opération par seconde par instance par service
- 1 Bench de montée en charge pour chaque service individuellement
 - Objectif : Valider les hypothèses de dimensionnement dans le contexte multi-instance. Mesurer l'écart de performance en multi-instance.
- 1 Bench de tenu en charge multi-service
 - Objectif : Valider les hypothèses de dimensionnement dans le contexte multi-service. Mesurer l'écart de performance en utilisation des services en //.

Les enjeux identifiés pour la mise en place de test de performance sur le GAR sont les suivants :

- Rassurer sur la montée en charge de l'application ;
- Valider les performances et l'endurance de l'application ;
- Anticiper les modifications du dimensionnement de l'infrastructure ;
- Identifier les points de contention restant.


2.2 Déroulement des tests de performances

2.2.1 Equipes Worldline

Pour réaliser cette mission l'équipe s'appuie sur les outils  et  qui permettent entre autres :

- Définition de scénarii représentatifs des cas d'utilisation;
- Simulation de différents niveaux de charge applicative ;
- Analyse fine des métriques aux différentes étapes des scénarii.

Cet outil est hébergé dans une infrastructure dédiée aux tests est géré par les équipes d'experts Worldline dans l'entité SDCO (Software Development Community Office).

Dans cette entité nous avons sollicité les experts applicatifs qui utilise l'outil  qui permet d'obtenir des analyses détaillées de la performance des applicatifs testés, ainsi :

- Détails des exécutions des différentes parties de l'application;
- Analyse du comportement de la JVM dans le traitement des objets;
- Analyse du comportement des composants attachés aux applications (BdD).

2.2.2 Applications

- Version majeure GAR: 7.1.90 des services :
 - Brique de collecte des données ENT
 - Brique d'import des données ENT
 - IHM d'affectations :
 - WS listes ressources
 - Accès Ressources
 - Batch de pré-affectations

2.3 Bilan des résultats des tirs multi instances

Services	Hypothèses	Résultats	Nombre d'instances Utilisées / théoriques pour valider l'hypothèse	Résultats obtenus	Comparatif production 2023/2020
IHM Affectation	1 000 responsables d'affectations réalisant des affectations en simultanée	Objectif atteint	4 / 4 théoriques pour valider l'hypothèse	Réaliser 5.37M affectations créées en 1h Créer 47 234 sessions en 1h	Equivalent au tir de la R5.2
WS Liste ressources	Gérer des pics de 100 requêtes par secondes	Objectif atteint	2 / 2 théoriques pour valider l'hypothèse	Gérer 500 requêtes par seconde	Equivalent au tir de la R5.2
Accès ressources	Gérer 50 000 accès ressources par minute	Objectif théoriquement atteint ¹	4 / 9 théoriques pour valider l'hypothèse	<u>Tir mono instance :</u> 6000 accès ressource par minute <u>Tir multi instance :</u> Avec 4 instances, réalisation de 27 000 accès ressources par minute Théoriquement, avec 9 instances, nous devrions atteindre 60 000 accès ressources par minutes	Equivalent au tir de la R5.2

Collecte ENT	Traiter les archives de 30 ENT (20 moyens et 10 petits) dans les mêmes temps que la R5.2, soit : ~1550 éléments/s pour les complets ~1500 éléments/s pour les deltas	Objectif atteint	4 / 4	<u>En mono instance :</u> 1413 éléments/s pour les complets 1735 éléments/s pour les deltas <u>En multi instance :</u> 3656 éléments/s pour les complets 2737 éléments/s pour les deltas	
Import ENT	Traiter les archives de 30 ENT (20 moyens et 10 petits) dans les mêmes temps que la R5.2, soit : ~2250 éléments/s pour les complets ~1250 éléments/s pour les deltas	Objectif atteint	4 / 4	<u>En mono instance :</u> 1966 éléments/s pour les complets 588 éléments/s pour les deltas <u>En multi instance (4 VM) :</u> 5482 éléments/s pour les complets 2152 éléments/s pour les deltas	Le multi instance permet de pallier la baisse des performances du batch d'import
Batch de pré-affectations	Gérer un pic de 12M de pré-affectations établissements Gérer un pic de 48 M de pré-affectations NE Gérer un pic de 840 000 d'affectations automatiques Gérer un pic de 720 000 de pré-affectations groupe	Objectif théoriquement atteint	1 / 22 théoriques pour valider l'hypothèse	Test effectué en mono instance uniquement à ce stade : 439 524 affectations automatiques en 4h41 2 856 867 de pré-affectations NE en 2h28 2 624 290 de pré-affectations établissements en 2h15	Générer 12 M de pré-affectations groupe en 0h55 Générer 48 M de pré-affectations NE en 10h Générer 840 000 affectations automatiques en 1h14 720 000 de pré-affectations groupe Générer 720 000 de pré-affectations groupe en 1h56

Néoload. Cependant les tests ont démontré que la plateforme peut théoriquement répondre à l'hypothèse (cf. ¹ L'objectif n'a pas pu être atteint du fait des limites de la plateforme §**Erreur ! Source du renvoi introuvable.**)

3 Présentation des résultats des tirs de performance

3.1 IHM Affectation

3.1.1 Description du tir

Le tir utilise le service suivant :

- L'IHM d'affectation

Le tir exécute 2 scénarios en parallèle :

- Un scénario d'affectation individuelle dont voici les différentes étapes :
 - Connexion à l'IHM grâce à l'IDP de test
 - Affichage de la liste des établissements
 - Affichage de la liste des ressources pour un établissement
 - Choix d'une ressource avec abonnement INDIV, affichage de la liste des élèves
 - Recherche des individus sans affectation
 - Sélection des individus unitairement
 - Attribution des ressources aux élèves
- Un scénario d'affectation établissement dont voici les différentes étapes :
 - Connexion à l'IHM grâce à l'IDP de test
 - Affichage de la liste des établissements
 - Affichage de la liste des ressources pour un établissement
 - Choix d'une ressource avec abonnement ETABL, affichage de la liste des populations
 - Sélection des populations unitairement
 - Attribution des ressources

La répartition définie pour le tir est la suivante :

- 70% pour le scénario d'affectation individuelle
- 20% pour le scénario d'affectation établissement
- 10% pour le scénario récupération des licences

Les seuils définis pour ce service sont les suivants :

- Performance: Temps de réponse < 3s (90 percentile)
- Disponibilité : Temps de réponse < 15s (99 percentile)

3.1.2 Stratégie des tirs

Trois tirs sont effectués avec les stratégies suivantes :

- Tir de montée en charge en mono instance
- Tir de performance en mono instance
- Tir de performance en multi instances

3.1.3 Optimisations

Une optimisation a été nécessaire au niveau du modèle objet Hibernate pour ne pas récupérer inopinément les adresses mail des accédants (qui ne sont pas utilisés dans le traitement)

3.1.4 Tirs mono instance

3.1.4.1 Configuration de la plateforme

Configuration	Valeur
Nombre de services déployés	1 service
Taille JVM	10 240 Mo

3.1.4.2 Tir de montée en charge en mono instance

Ce tir n'a pas eu besoin d'être réalisé dans le cadre de la 7.1.90

3.1.4.3 Tir de performance en mono instance

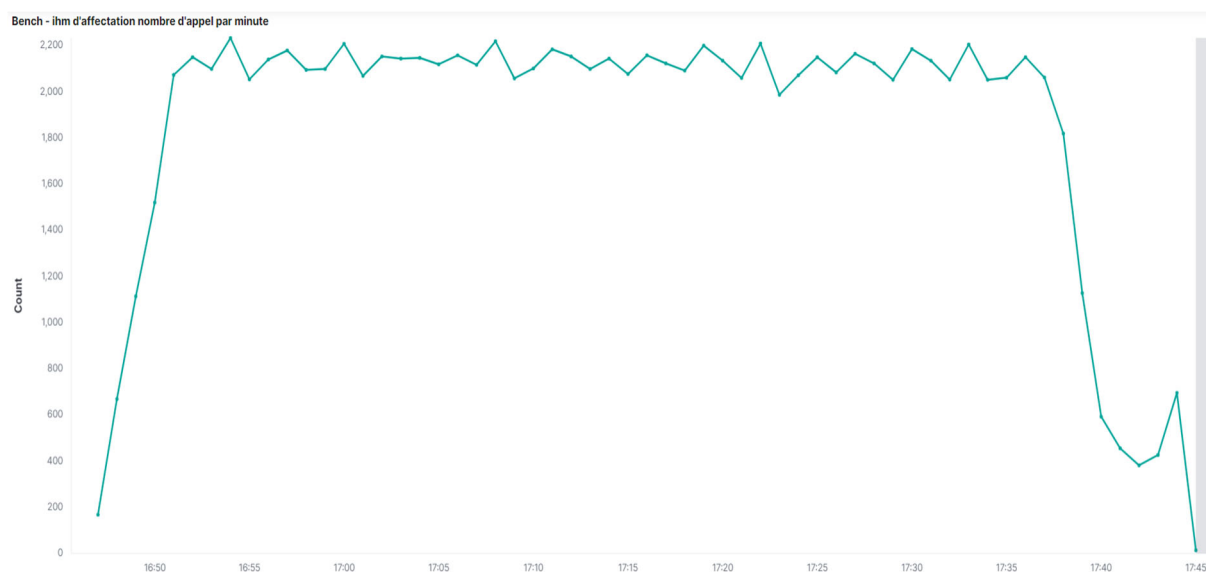
3.1.4.3.1 Informations sur le tir

Configuration	Valeur
Durée du tir	1 heures
Nombre de responsables d'affectations	350

Configuration de la plateforme d'injection Néoload :

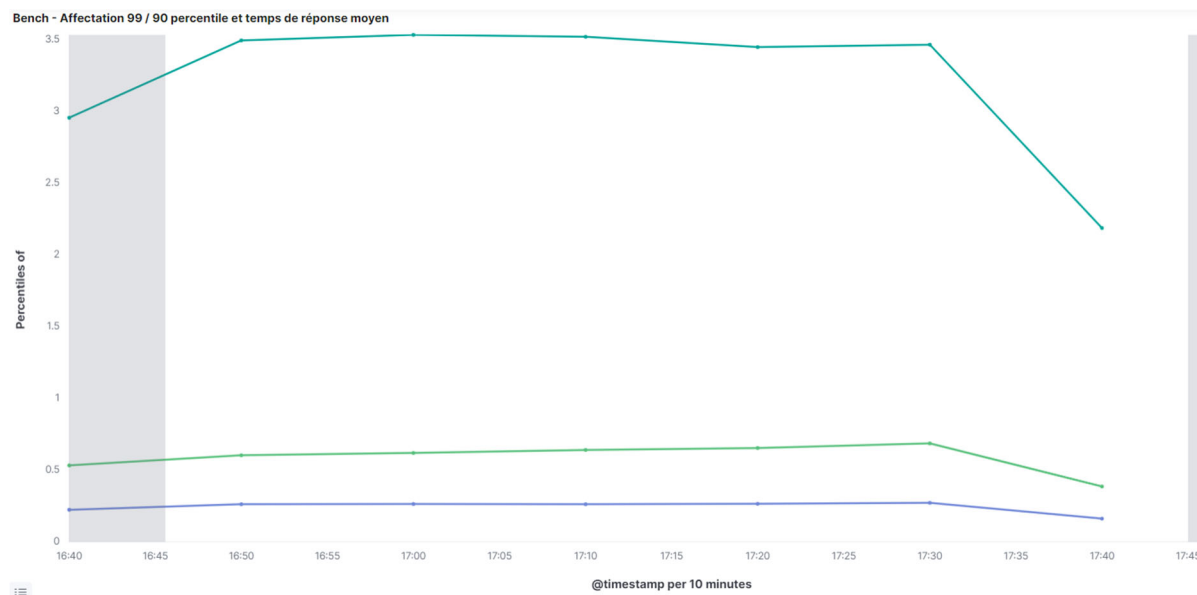
- Nombre initial d'utilisateurs : 1
- Incréments de 1 utilisateur
- Toutes les 1 secondes
- Nombre maximum d'utilisateurs : 350
- Limite attente au bout de 5 minutes et 39 secondes

3.1.4.3.2 Nombre d'appels à l'IHM d'affectation par minute



Le service se stabilise à une moyenne de 2232 appels par minute.

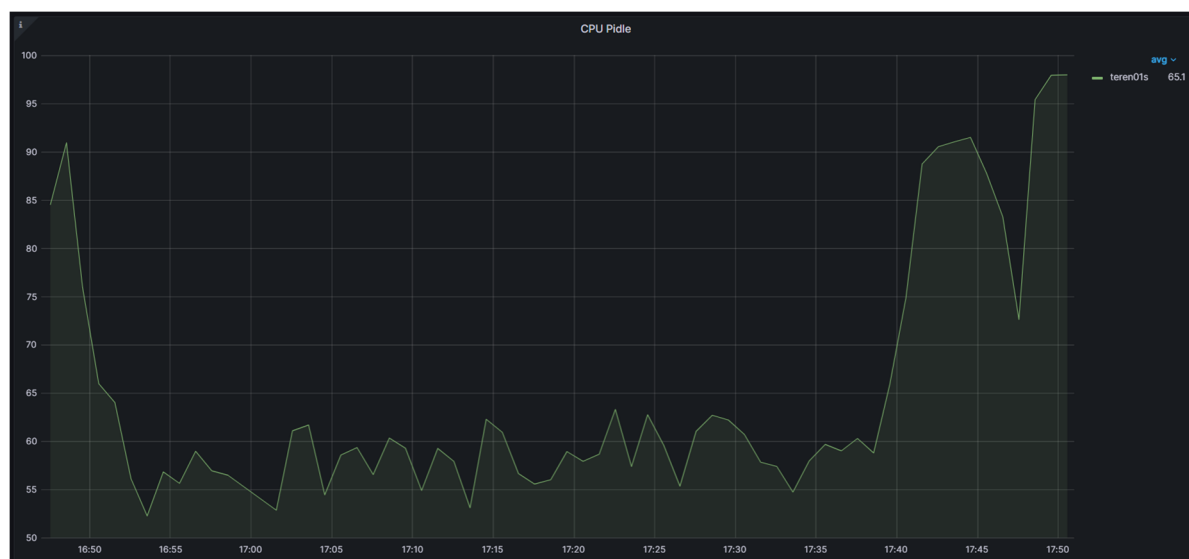
3.1.4.3.3 Temps de réponse moyen / 90 percentiles / 99 percentiles



Le service se stabilise à une moyenne de :

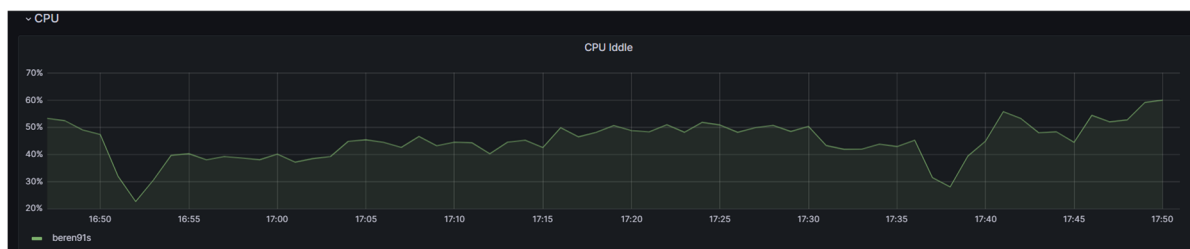
- 262 ms pour le temps de réponse moyen
- 640 pour les 90 percentiles
- 3 secondes et 535 ms pour les 99 percentiles

3.1.4.3.4 Utilisation CPU machine applicative



La charge CPU sur le serveur applicatif est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.1.4.3.5 Utilisation CPU machines BDD



La charge CPU sur le serveur BDD est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.1.1 Conclusion tirs mono instance

Configuration	Valeur
Nombre d'appels par minute	Entre 1 986 et 2 232 appels Moyenne : 2 109 appels
Temps de réponse (moyenne)	Entre 180 et 345 ms Moyenne : 240 ms
Temps de réponse (90 Percentiles)	Entre 1 s 5ms et 1 s 19 ms Moyenne : 1 s 10 ms
Temps de réponse (99 Percentiles)	Entre 2 s 230 ms et 4 s 378 ms Moyenne : 2s 470ms
Utilisation du CPU machine applicative	45 % d'utilisation du CPU en moyenne 55 % d'utilisation du CPU maximum
Utilisation du CPU machines BDD	6 % d'utilisation du CPU en moyenne 7 % d'utilisation du CPU maximum
Nombre d'affectations créées	3 481 946 Affectations
Nombre de sessions de responsables d'affectations en parallèle	350
Nombre de sessions total	16 837

3.1.2 Tirs multi instances

3.1.2.1 Objectif du tir multi instances

Voici la liste des objectifs de ce tir :

- Avoir 1 000 responsables d'affectations réalisant des affectations en simultanée
- Respect des seuils de :
 - Performance: Temps de réponse < 3s (90 percentile)
 - Disponibilité : Temps de réponse < 15s (99 percentile)

3.1.2.2 Configuration de la plateforme

En prenant en compte les résultats du tir mon instance : 1 instance permet de gérer 350 responsables d'affectations, pour gérer 1000 responsables d'affectations il faudra 3 instances + 1. Voici la configuration mise en place pour ce service pour répondre aux objectifs.

Configuration	Valeur
---------------	--------

Nombre de services déployés	4 services
Taille JVM	10 240 Mo

Configuration de la plateforme d'injection Néoload :

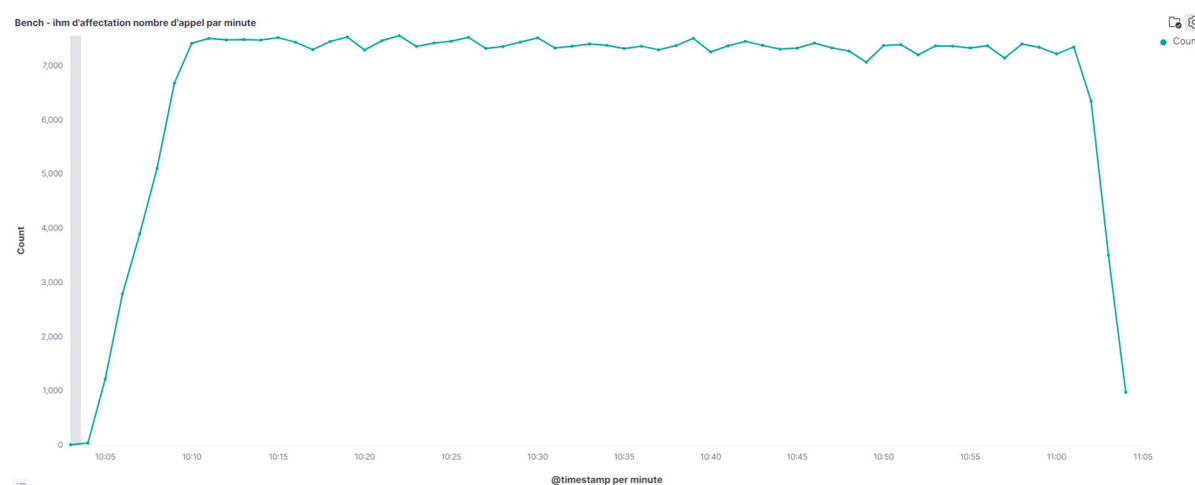
- Nombre initial d'utilisateurs : 1
- Incréments de 1 utilisateur
- Toutes les 1 secondes
- Nombre maximum d'utilisateurs : 1000
- Limite attente au bout de 10 minutes et 10 secondes

3.1.2.3 Tir de performance en multi instance

3.1.2.3.1 Informations sur le tir

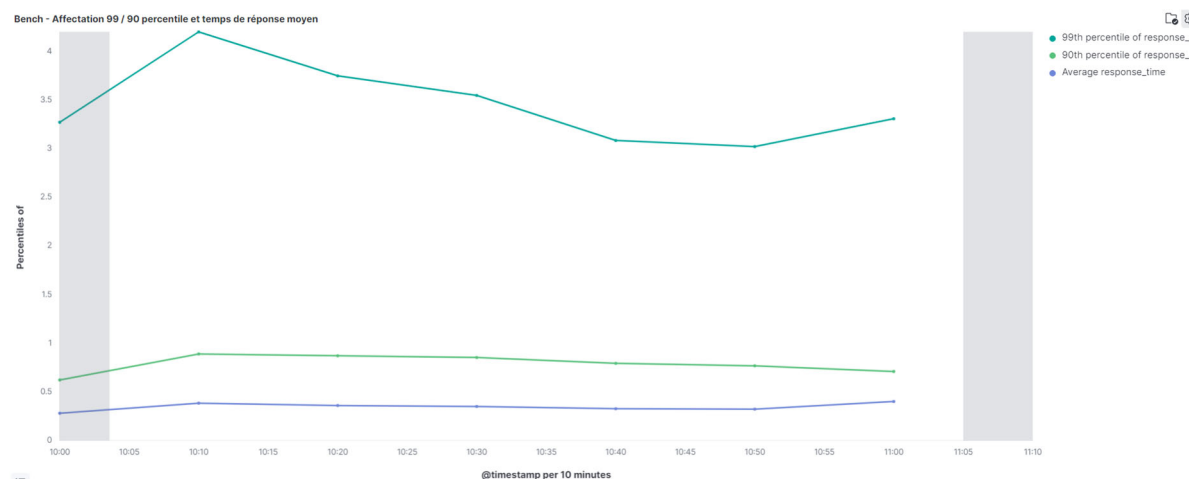
Configuration	Valeur
Durée du tir	1 heures
Nombre de responsables d'affectations	1 000

3.1.2.3.2 Nombre d'appels à l'IHM d'affectation par minute



Le service se stabilise à une moyenne de 7 523 appels par minute.

3.1.2.3.3 Temps de réponse moyen / 90 percentiles / 99 percentiles

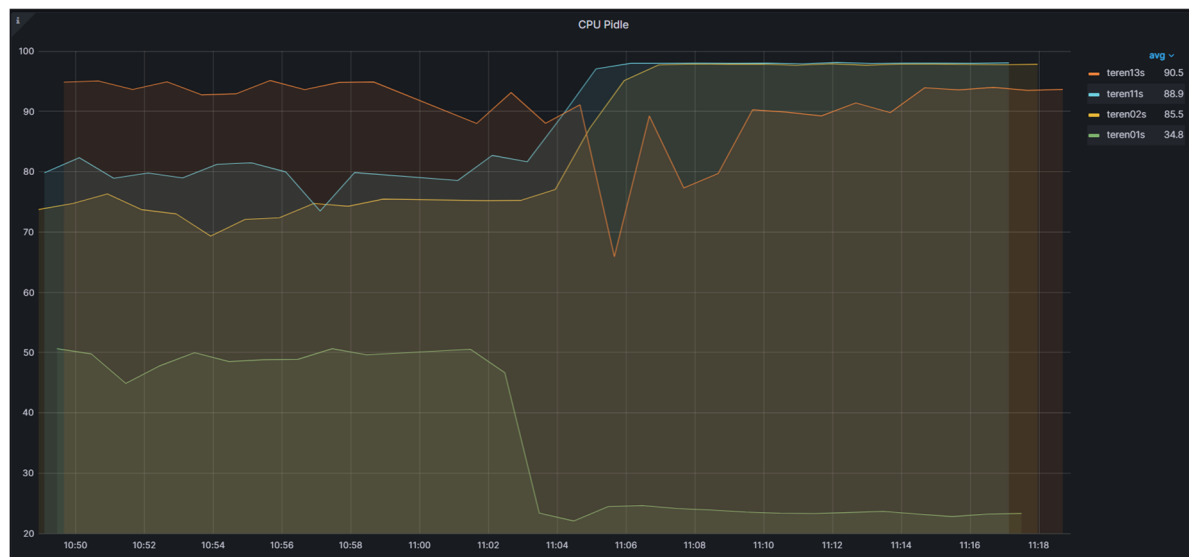


Le service se stabilise à une moyenne de :

- 338 ms pour le temps de réponse moyen
- 798 ms pour les 90 percentiles
- 3 secondes et 330 ms pour les 99 percentiles

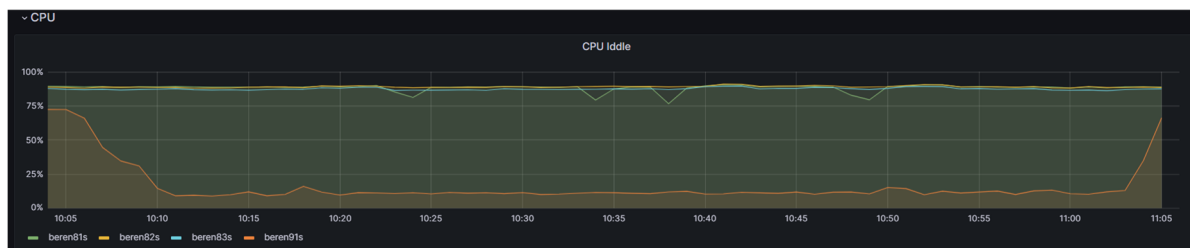
Nous constatons un pic de temps de réponse pour 99 percentiles à 4s 196ms, mais ce pic reste en dessous du seuil de disponibilité.

3.1.2.3.4 Utilisation CPU machines applicatives



La charge CPU sur les serveurs applicatifs est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.1.2.3.5 Utilisation CPU machines BDD



La charge CPU sur les serveurs BDD est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.1.3 Conclusion tirs multi instances

Configuration	Valeur
Nombre d'appels par minute	Entre 7 069 et 7 523 Moyenne : 7 296
Temps de réponse (moyenne)	Entre 281 ms et 384 ms Moyenne : 338 ms
Temps de réponse (90 Percentiles)	Entre 622 ms et 890 ms Moyenne : 756 ms
Temps de réponse (99 Percentiles)	Entre 3s 270ms et 4s 196s Moyenne : 3 s 733 ms
Utilisation du CPU machines applicatives	40 % d'utilisation du CPU en moyenne 89 % d'utilisation du CPU maximum
Utilisation du CPU BDD	17 % d'utilisation du CPU en moyenne 20 % d'utilisation du CPU maximum
Nombre d'affectations créées	5 369 973 Affectations
Nombre de sessions de responsables d'affectations en parallèle	1 000
Nombre de sessions total	47 234

Ce tir nous permet de valider que 4 instances de ce service permettent de répondre aux hypothèses du tir souhaité :

- 1000 responsables d'affectations réalisant des affectations en parallèle.
- Respect des seuils de performance et disponibilité

3.1.4 Focus sur la volumétrie de production de l'année 2019/2020

Concernant l'utilisation de la plateforme de production sur l'année 2019/2020 avec une volumétrie de 3.5M d'accédants, le service a réalisé :

- 1,33 Millions d'affectations créées par jour
- 5900 sessions créées par jour

Ce test de performance nous permet de confirmer que 4 instances de ce service permettent de:

- Réaliser 5 369 973 affectations créées (en 1h) : ce qui représente environ 4 fois le volume de production d'une journée.
- De créer 47 234 sessions (en 1h) : ce qui représente environ 8 fois le volume de production d'une journée.

3.2 WS Liste ressources

3.2.1 Description du tir

Le tir utilise le service suivant :

- Le web service de liste ressources

Le scénario est un appel au verbe du web service pour lister les ressources affectées pour un accédant.

Hypothèse du tir :

Chaque accédant a été affecté à 15 ressources.

Les seuils définis pour ce service sont les suivants :

- Performance: Temps de réponse < 0.5s (90 percentile)
- Disponibilité : Temps de réponse < 5s (99 percentile)

3.2.2 Stratégie des tirs

Un tir est effectuée avec les stratégies suivantes :

- Tir de performance en multi instances

3.2.1 Optimisations

Il n'a pas été nécessaire de réaliser d'optimisations pour ce service.

3.2.2 Tirs multi instances

3.2.2.1 Objectif du tir multi instances

Voici la liste des objectifs de ce tir :

- Gérer des pics de 30 000 appels par minute
- Respect des seuils de :
 - Performance: Temps de réponse < 0.5s (90 percentile)
 - Disponibilité : Temps de réponse < 5s (99 percentile)

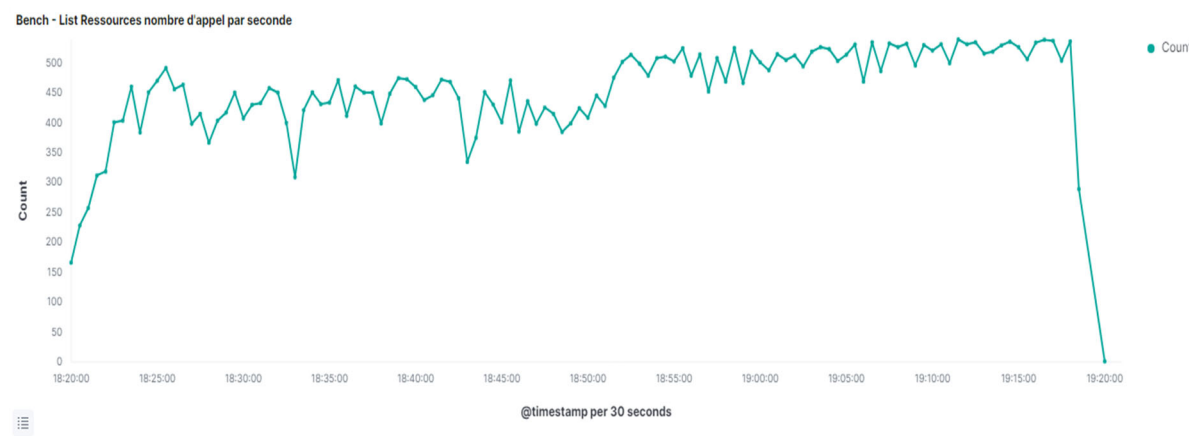
3.2.2.2 Configuration de la plateforme

En prenant en compte les résultats du tir mon instance : 1 instance permet de gérer des pics à 15 000 appels à la minute, pour gérer des pics à 30 000 appels à la minute il faudra 2 instances. Voici la configuration mise en place pour ce service pour répondre aux objectifs.

Configuration	Valeur
Nombre de services déployés	2 services
Taille JVM	2048 Mo

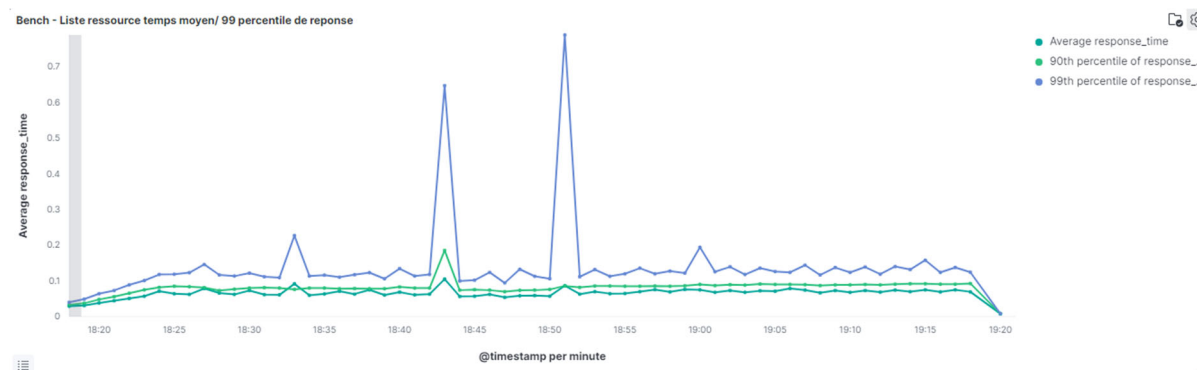
3.2.2.3 Tir de performance en multi instance

3.2.2.3.1 Nombre d'appels au liste ressources par seconde



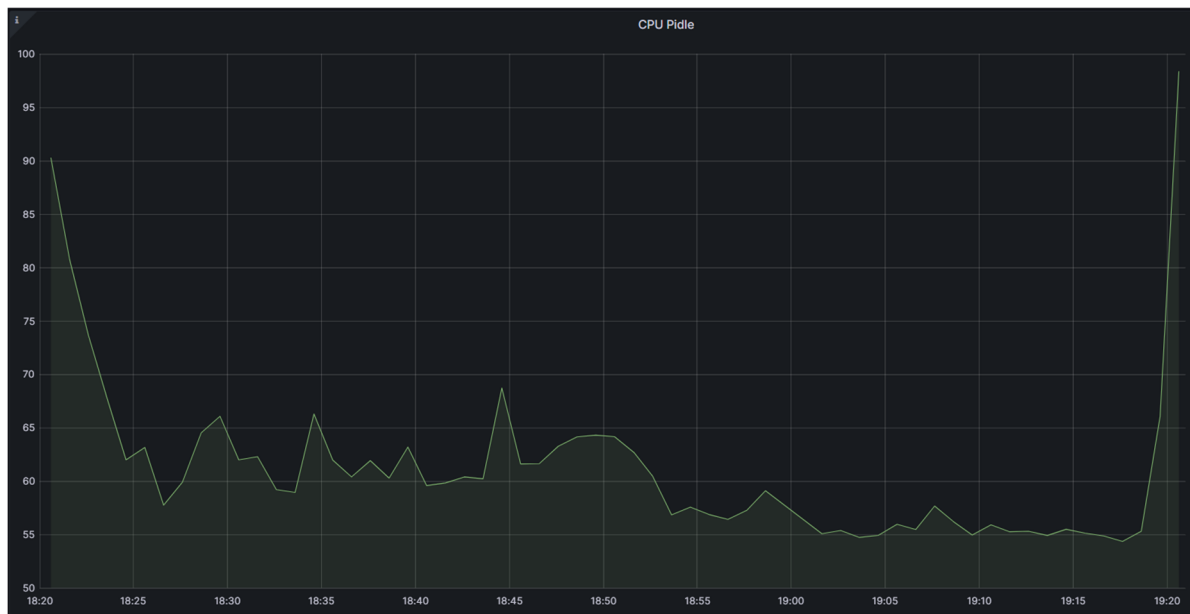
Le service se stabilise à une moyenne de 520 appels par seconde au WS liste ressources.

3.2.2.3.2 Temps de réponse moyen / 90 percentiles / 99 percentiles



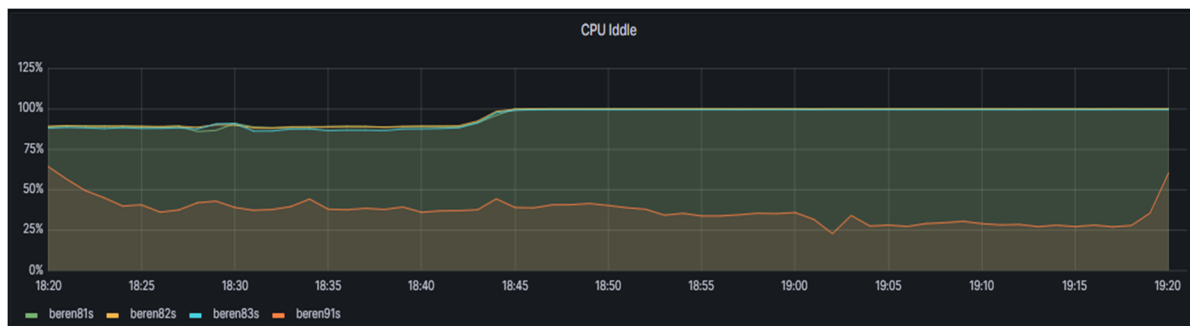
Le service se stabilise à une moyenne de 76 ms pour le temps de réponse moyen.
 Les pics de temps de réponses restent sous les seuils de performance et de disponibilité.

3.2.2.3.3 Utilisation CPU machines applicatives



La charge CPU sur les serveurs applicatifs est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.2.2.3.4 Utilisation CPU machines BDD



La charge CPU sur les serveurs BDD est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.2.3 Conclusion tirs multi instances

Configuration	Valeur
Nombre d'appels par seconde	Entre 465 et 535 Moyenne 500 appels
Temps de réponse (moyenne)	Entre 29ms et 105ms Moyenne 67 ms
Temps de réponse (90 Percentiles)	Entre 33ms et 186ms Moyenne 109 ms

Temps de réponse (99 Percentiles)	Entre 118ms et 790ms Moyenne 468 ms
Utilisation du CPU machine applicative	40 % d'utilisation du CPU en moyenne 55 % d'utilisation du CPU maximum
Utilisation du CPU BDD	8 % d'utilisation du CPU en moyenne 10 % d'utilisation du CPU maximum

Ce tir nous permet de valider que 2 instances de ce service permettent de répondre aux hypothèses du tir souhaité :

- Gestion d'un pic de 30 000 appels par minute
- Respect des seuils de performance et disponibilité

3.3 Accès ressources

3.3.1 Description du tir

Le tir réalise un accès à ressources en utilisant les services suivant :

- Le service d'accès aux ressources
- Le simulateur d'IDP
- Le simulateur de ressources (CAS et SAML)

Le tir répartie les accès aux ressources avec le ratio suivant :

- **80%** des appels sur une ressource CAS
- **20%** des appels sur une ressource SAML

Les seuils définis pour ce service sont les suivants :

- Performance: Temps de réponse < 1.5s (90 percentile)
- Disponibilité : Temps de réponse < 15s (99 percentile)

3.3.2 Stratégie du tir

Trois tirs sont effectués avec les stratégies suivantes :

- tir de montée en charge
- tir de performance mono instance
- tir de performance multi instances

3.3.3 Optimisations

Il n'a pas été nécessaire de réaliser d'optimisations pour ce service.

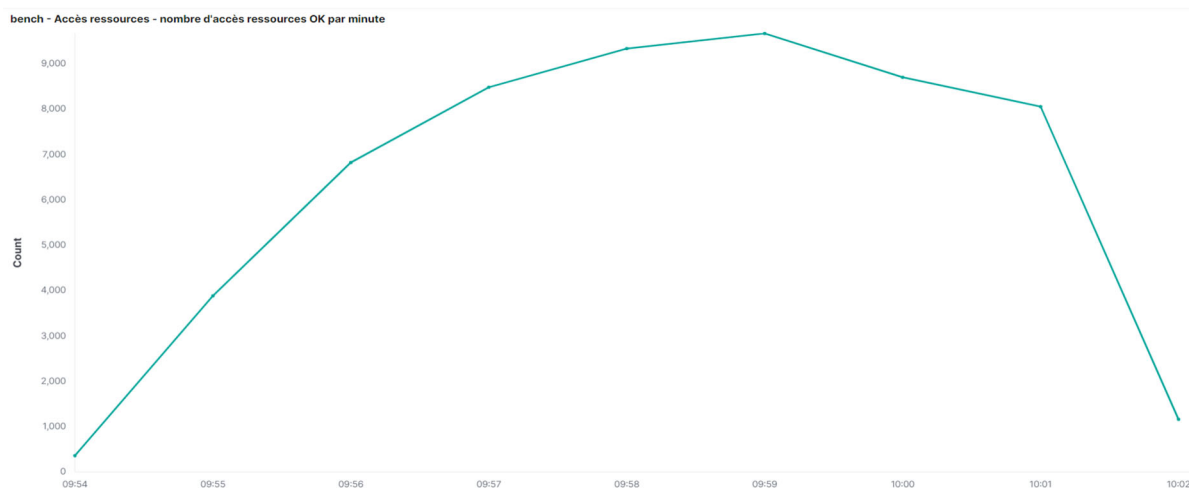
3.3.4 Tirs mono instance

3.3.4.1 Configuration de la plateforme

Configuration	Valeur
Nombre de services déployés	1 service
Taille JVM	8192 Mo

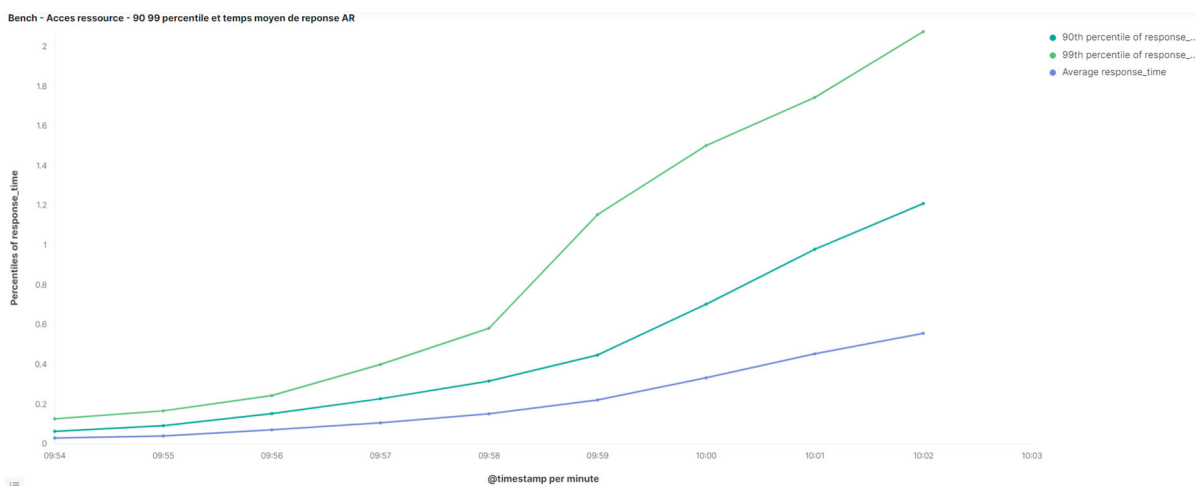
3.3.4.2 Tir de montée en charge en mono instance

3.3.4.2.1 Nombre d'accès ressources par minute



Lors de ce tir, nous avons réussi à obtenir un pic à 9 672 accès ressources par minute.

3.3.4.2.2 Temps de réponse moyen / 90 percentiles / 99 percentiles



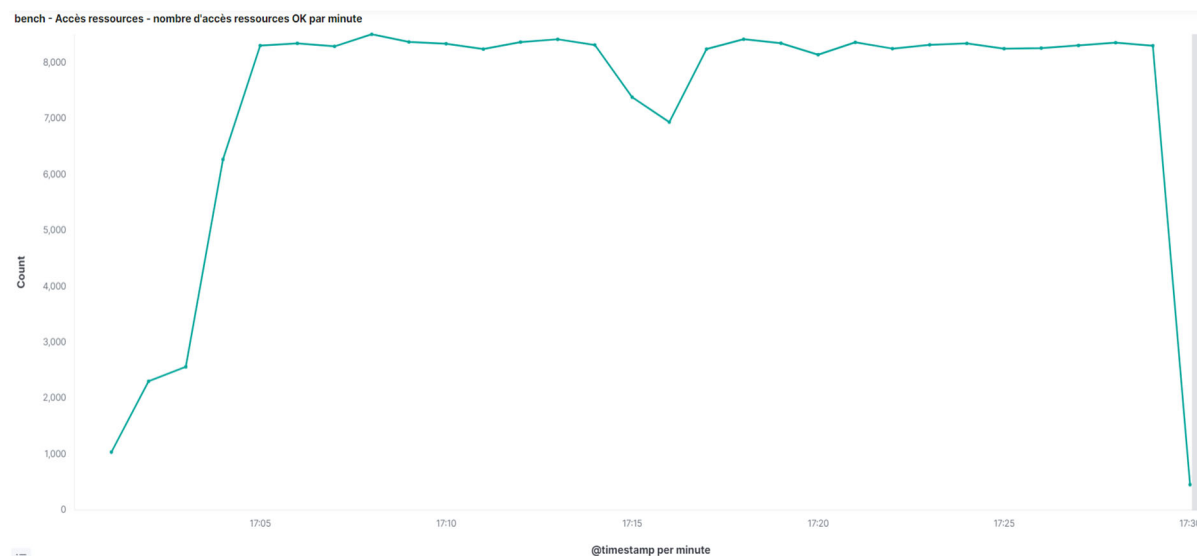
Lors de ce tir, nous identifions une augmentation significative des temps de réponse au bout de 5 minutes avec dépassement du seuil de performance.

3.3.4.2.3 Conclusion

Un seul service peut accepter une moyenne de 9 600 appels par minute. Au-dessus de ce seuil, le nombre d'appels augmente légèrement mais impacte fortement les temps de réponse du service d'accès ressource.

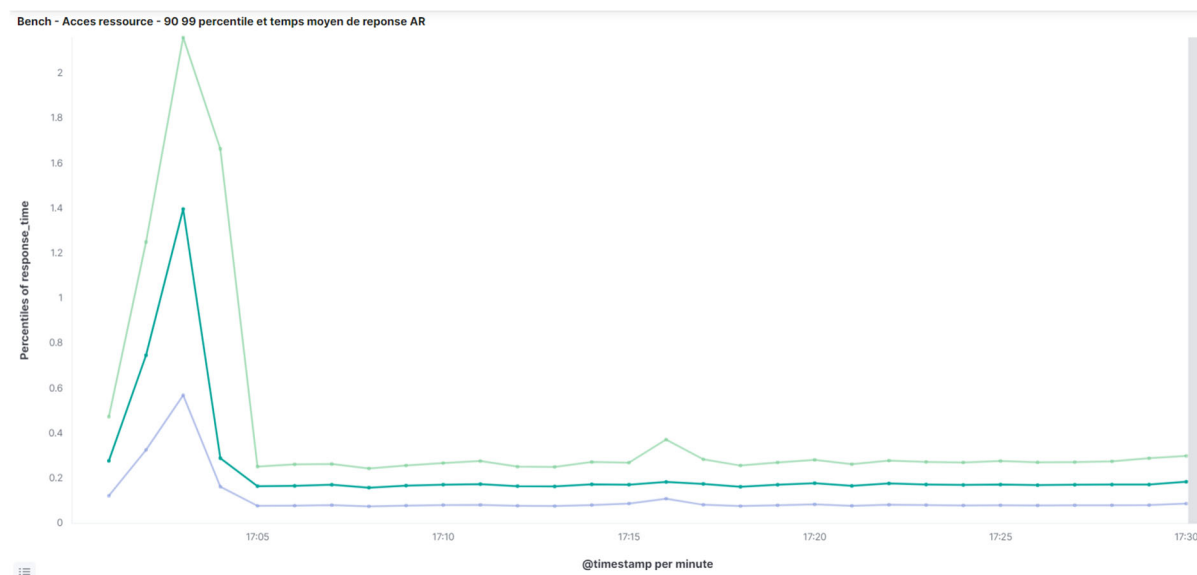
3.3.4.3 Tir de performance en mono instance

3.3.4.3.1 Nombre d'accès ressources par minute



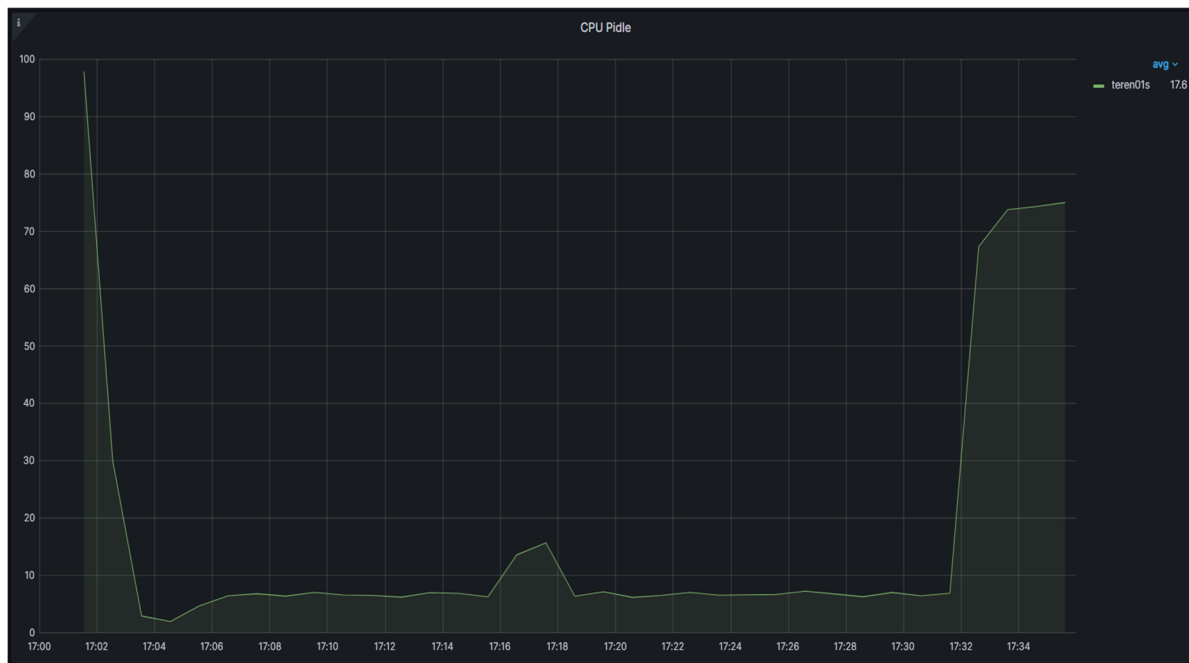
Le service se stabilise à une moyenne de 8 500 appels par minute au WS liste ressources.

3.3.4.3.2 Temps de réponse moyen / 90 percentiles / 99 percentiles



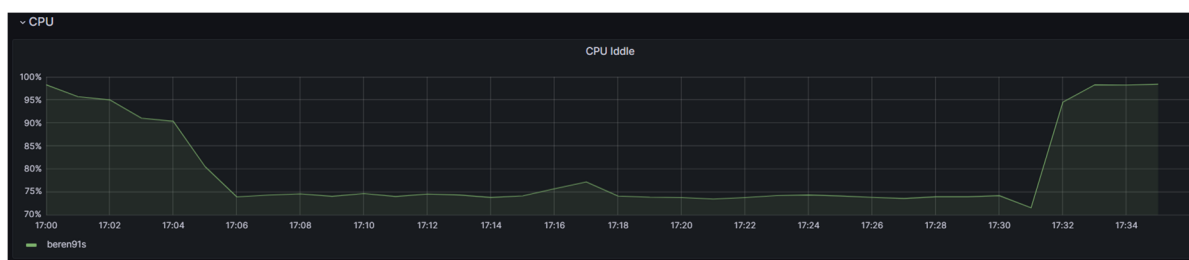
Le service se stabilise à une moyenne de 280 ms pour le temps de réponse moyen.

3.3.4.3.3 Utilisation CPU machine applicative



La charge CPU sur le serveur applicatif est au-dessus des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir mais ne dégrade pas le service en lui-même.

3.3.4.3.4 Utilisation CPU machines BDD



La charge CPU sur les serveurs BDD est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.3.5 Conclusion tirs mono instance

Configuration	Valeur
Nombre d'accès ressource par minute	Entre 7 000 et 8 600 appels
Temps de réponse (moyenne)	Entre 110 ms et 590 ms
Temps de réponse (90 Percentiles)	Entre 200 ms et 1s 400 ms
Temps de réponse (99 Percentiles)	Entre 390 ms et 2s 500 ms
Utilisation du CPU machine applicative	90 % d'utilisation du CPU en moyenne 95 % d'utilisation du CPU maximum
Utilisation du CPU machines BDD	2.5 % d'utilisation du CPU en moyenne 3 % d'utilisation du CPU maximum

3.3.6 Tirs multi instances

3.3.6.1 Objectif du tir multi instances

Voici la liste des objectifs de ce tir :

- Gérer des pics de 100 000 appels par minute
- Respect des seuils de :
 - Performance: Temps de réponse < 0.5s (90 percentile)
 - Disponibilité : Temps de réponse < 5s (99 percentile)

3.3.6.2 Configuration de la plateforme

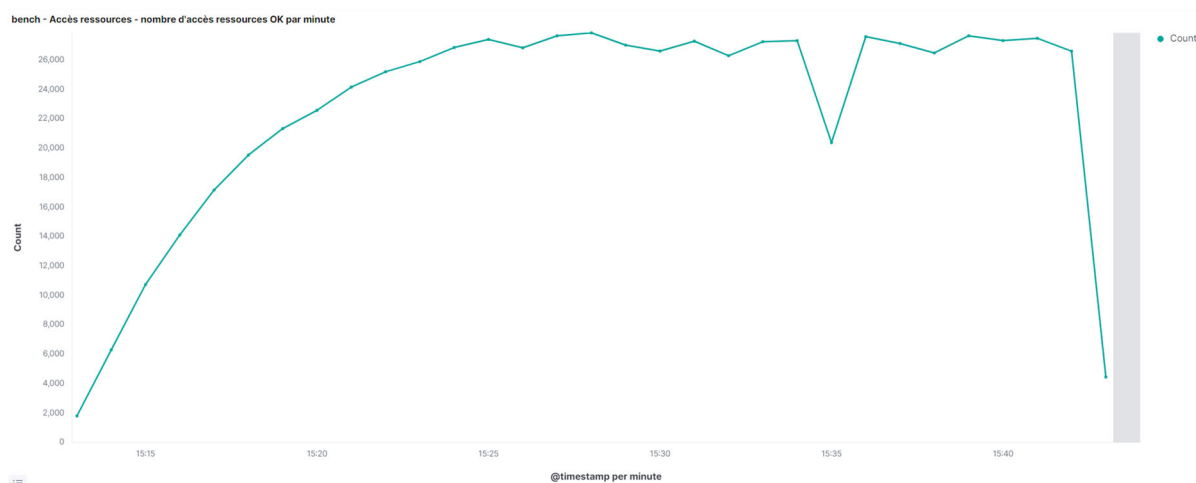
En prenant en compte les résultats du tir mon instance : 1 instance permet de gérer des pics à 9 500 appels à la minute, pour gérer des pics à 100 000 appels à la minute il faudra 11 instances + 1. Comme nous avons rencontré des limitations sur notre plateforme Néoload, nous avons choisi de n'utiliser que 4 machines afin que le tir multi instances utilise pleinement chaque instance.

Voici la configuration mise en place pour ce service pour répondre aux objectifs.

Configuration	Valeur
Nombre de services déployés	4 services
Taille JVM	8192 Mo

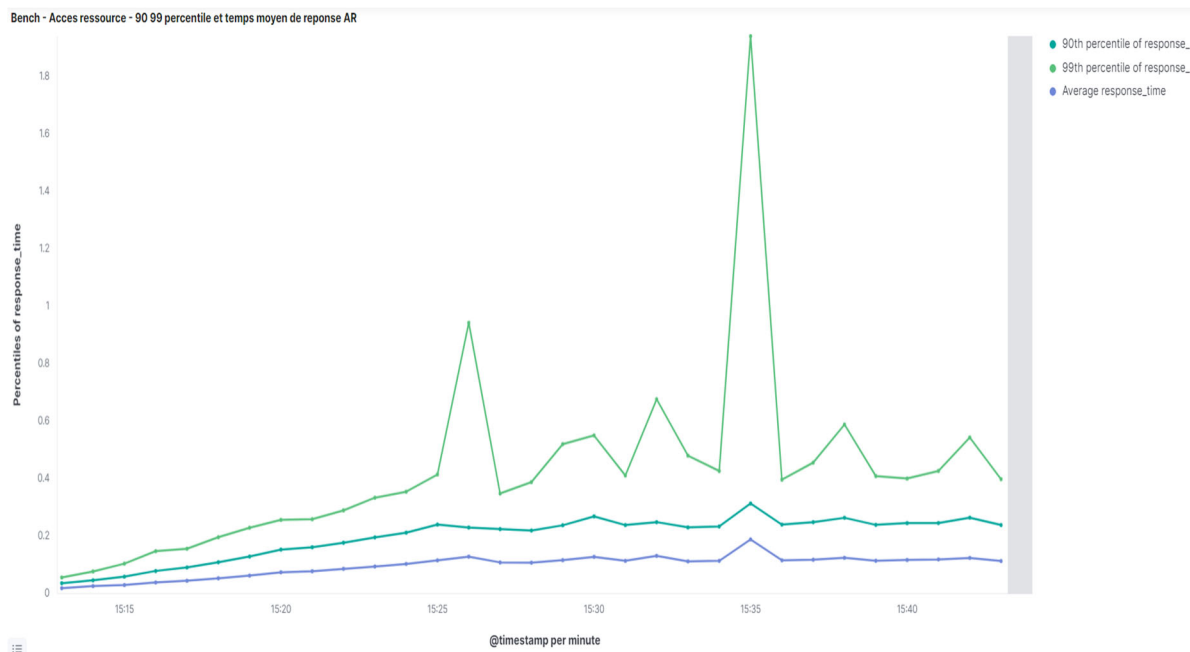
3.3.6.3 Tir de performance en multi instance

3.3.6.3.1 Nombre d'accès ressources par minute



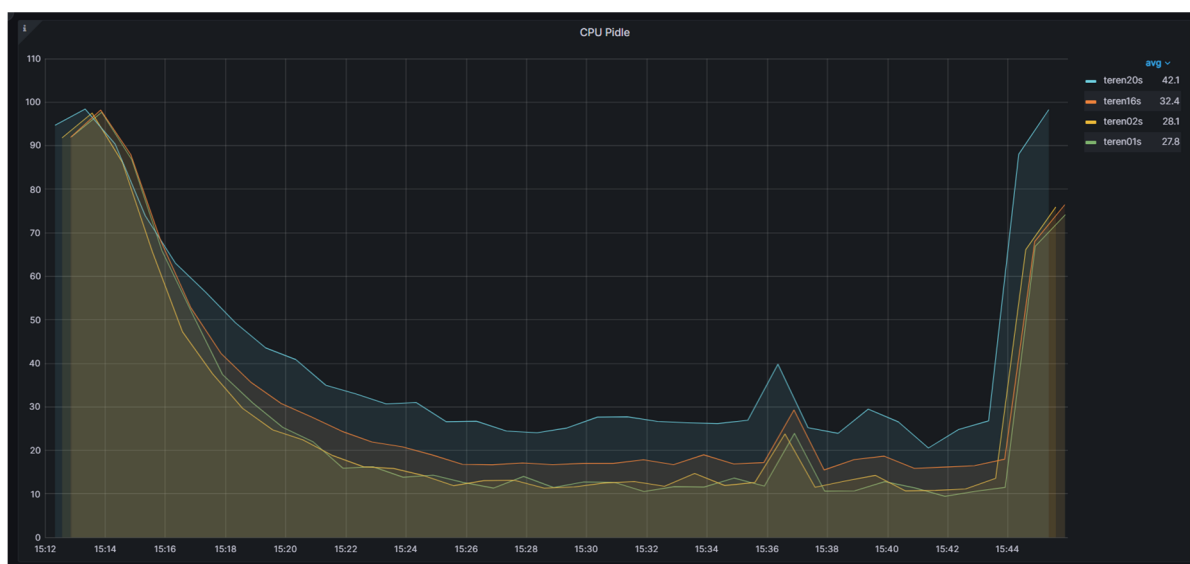
Lors de ce tir, nous avons réussi à obtenir un pic à 27 840 accès ressources par minute.

3.3.6.3.2 Temps de réponse moyen / 90 percentiles / 99 percentiles



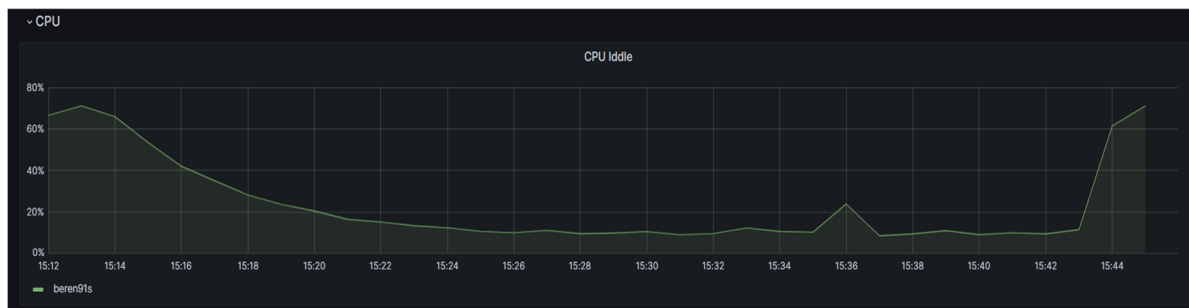
Le service se stabilise à une moyenne de 120 ms pour le temps de réponse moyen.
 Les pics de temps de réponses restent sous les seuils de performance et de disponibilité.

3.3.6.3.3 Utilisation CPU machines applicatives



La charge CPU sur les serveurs applicatifs est au-dessus des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir mais ne dégrade pas le service en lui-même.

3.3.6.3.4 Utilisation CPU machines BDD



La charge CPU sur les serveurs BDD est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.3.7 Conclusion tirs multi instances

Configuration	Valeur
Nombre d'appels par minute	27 000 appels Moyenne
Temps de réponse (moyenne)	Entre 20 et 189 ms Moyenne 96 ms
Temps de réponse (90 Percentiles)	Entre 36 et 313 ms Moyenne 174 ms
Temps de réponse (99 Percentiles)	Entre 56 et 1s 944 ms Moyenne 1s
Utilisation du CPU machine applicative	75 % d'utilisation du CPU en moyenne 90 % d'utilisation du CPU maximum
Utilisation du CPU BDD	10 % d'utilisation du CPU en moyenne 12 % d'utilisation du CPU maximum

Ce test de performance nous permet de valider directement que 4 instances de ce service permettent d'accéder à un pic de 27 000 appels/min. La limitation de notre plateforme d'injecteurs Néoload ne nous permet pas d'atteindre un pic à 50 000 appels par minute.

Cependant, une projection peut être théoriquement faite avec 9 VM sur les bases des résultats de 4VM : Nous devrions atteindre théoriquement 60 000 accès ressources par minute.

3.4 Collecte et import des données ENT

3.4.1 Description du tir

Le tir utilise les services suivant :

- Le service SFTP de dépôt d'archives des données ENT
- Le service de collecte des données ENT (KOSMOS)
- Le service d'import des données ENT

Les archives de type COMPLET sont déposées en entrée de la chaîne d'import des données ENT (SFTP) pour être traitées par le service de collecte puis par le service de brique d'import des données ENT.

Lors de ce tir, 3 scénarios seront exécutés :

- Scénario 1: Collecte et Import full d'un jeu de donnée anonymisé (nom/prénom/mail) de production *2 + X projets pour atteindre ~7M d'accédants sur 1 instance
- Scénario 2: Collecte et Import de ce jeu de donnée modifié à 100% sur 1 instance
- Scénario 3: Purge puis Collecte sur N instances et Import sur N instances de ce jeu de donnée

3.4.2 Stratégie du tir

La stratégie du tir se découpe en 3 étapes :

- Import d'un jeu de données complet (import FULL en mono instance)
- Import d'un jeu de données 100% modifié (import DELTA en mono instance)
- Purge et import du même jeu de données sur N instances (import FULL en multi instance)

3.4.3 Optimisations

Il n'a pas été nécessaire de réaliser d'optimisations pour ce service.

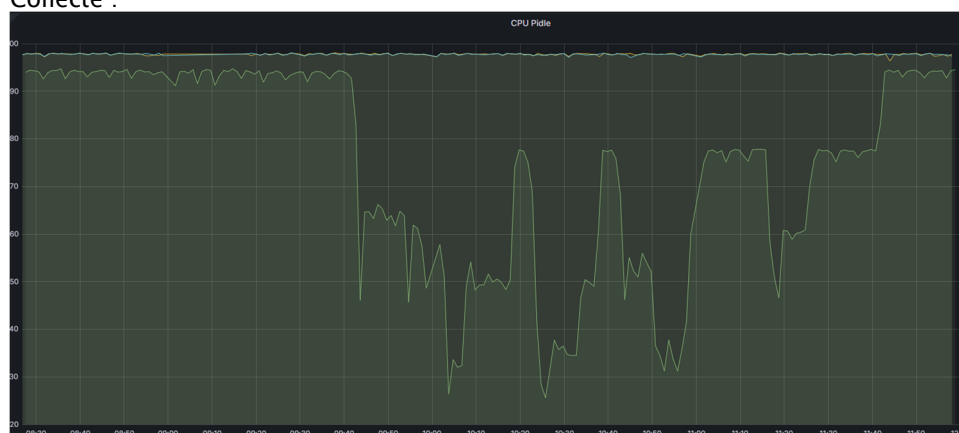
3.4.4 Tirs import FULL en mono instance

3.4.4.1 Configuration de la plateforme

Configuration	Valeur
Nombre de services collecte des données ENT déployés	1 service
Taille JVM	6 144 Mo
Nombre de services import des données ENT déployés	1 service
Taille JVM	10 240 Mo

3.4.4.2 Utilisation CPU machine applicative

Collecte :



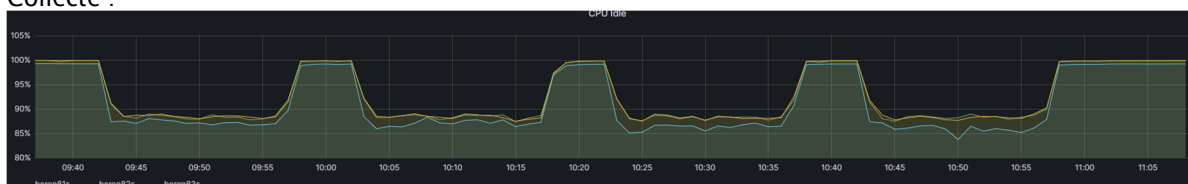
Import :



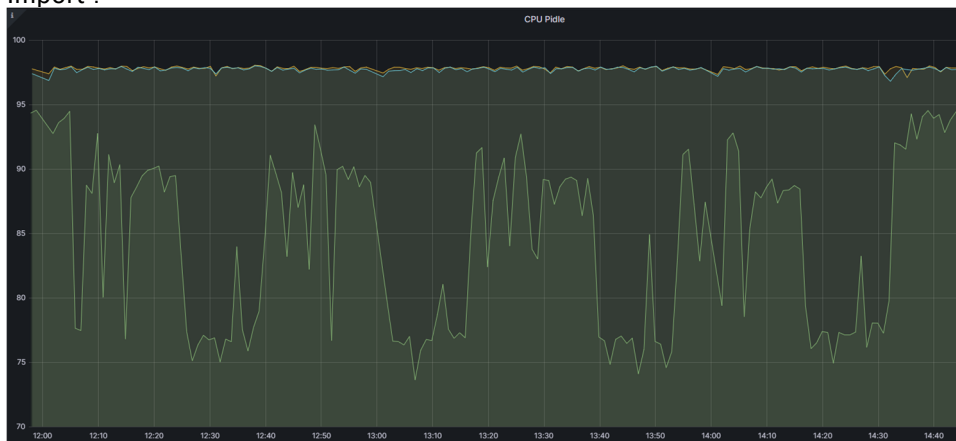
La charge CPU sur le serveur applicatif est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.4.4.3 Utilisation CPU machine applicative

Collecte :



Import :



3.4.4.4 Résultat du traitement

Configuration	Valeur
Nombre d'archives traités	4
Nombres d'éléments à traiter	17 222 220 éléments
Temps total du traitement de la collecte	3 heures 23 minutes
Nombre d'éléments traités par seconde pour la collecte	1413 éléments/seconde

Nombres d'éléments à traiter	17 222 220 éléments
Temps total du traitement de la brique d'import	2 heures et 25 minutes
Nombre d'éléments traités par seconde pour la brique d'import	1966 éléments/seconde
Utilisation CPU machine applicative	40 % d'utilisation du CPU en moyenne 70 % d'utilisation du CPU maximum
Utilisation CPU machine BDD	13 % d'utilisation du CPU en moyenne pour la collecte 15 % d'utilisation du CPU en moyenne pour l'import

3.4.5 Conclusion import FULL mono instance

La brique de collecte permet de gérer 1413 éléments/seconde et la brique d'import permet de gérer 1966 éléments/seconde.

En comparaison aux résultats des tirs de performance réalisés en 2020, La brique d'import des données ENT a une baisse de performances d'environ 13% sur le traitement des données.

3.4.1 Tirs import DELTA en mono instance

3.4.1.1 Configuration de la plateforme

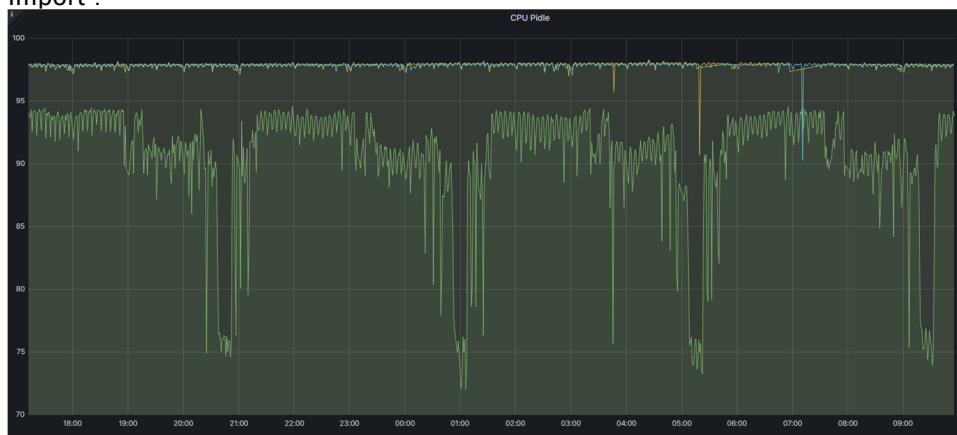
Configuration	Valeur
Nombre de services collecte des données ENT déployés	1 service
Taille JVM	6 144 Mo
Nombre de services import des données ENT déployés	1 service
Taille JVM	10 240 Mo

3.4.1.2 Utilisation CPU machine Applicative

Collecte :

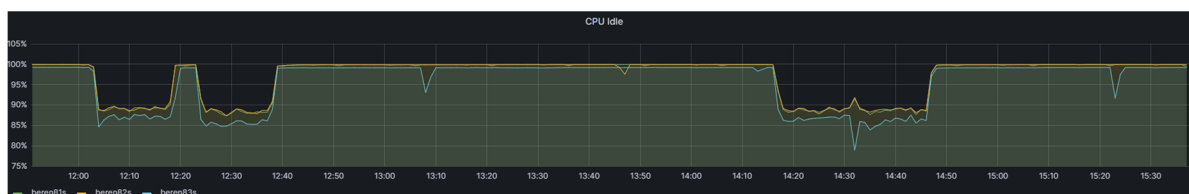


Import :

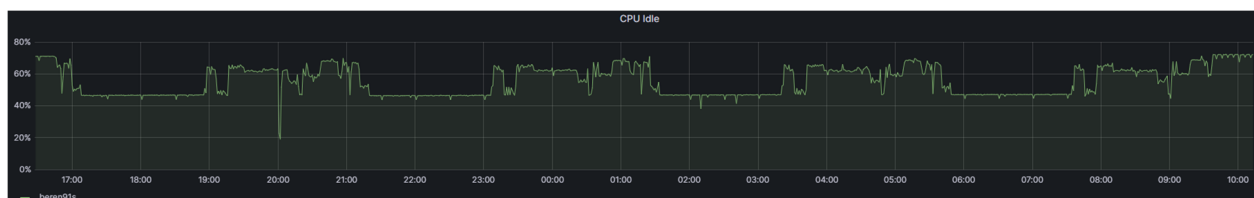


3.4.1.3 Utilisation CPU machine BDD

Collecte :



Import :



3.4.1.4 Résultat du traitement

Configuration	Valeur
Nombre d'archives traités	4
Nombres d'éléments à traiter	35 494 728 éléments
Temps total du traitement de la collecte	5 heures 40 minutes
Nombre d'éléments traités par seconde pour la collecte	1735 éléments/seconde
Nombres d'éléments à traiter	35 494 728 éléments
Temps total du traitement de la brique d'import	16 heures et 46 minutes
Nombre d'éléments traités par seconde pour la brique d'import	588 éléments/seconde
Utilisation CPU machine applicative	25 % d'utilisation du CPU en moyenne 60 % d'utilisation du CPU maximum

Utilisation CPU machine BDD	13 % d'utilisation du CPU en moyenne pour la collecte 15 % d'utilisation du CPU en moyenne pour l'import
------------------------------------	---

3.4.2 Conclusion import DELTA mono instance

La brique de collecte permet de gérer 1735 éléments/seconde et la brique d'import permet de gérer 588 éléments/seconde.

En comparaison aux résultats des tirs de performance réalisés en 2020, La brique d'import des données ENT a une baisse de performances divisé par 2 à la suite de la refonte du batch d'import mais peut être palier par la mise en place du multi instance.

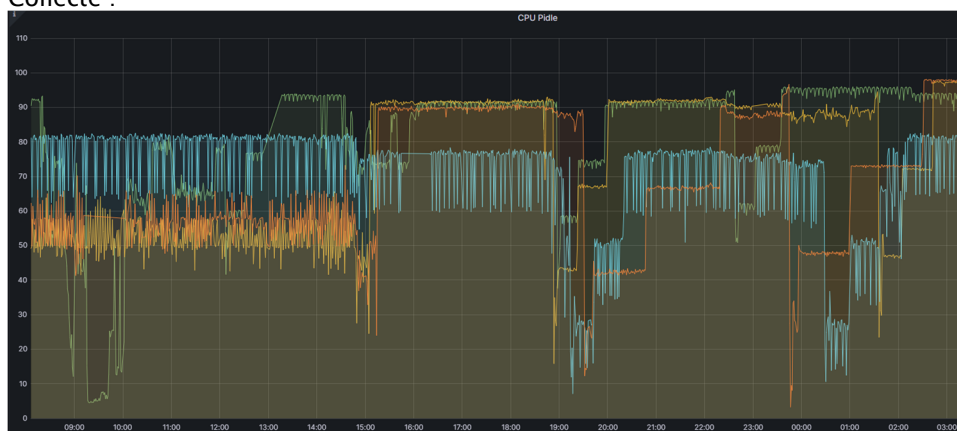
3.4.1 Tirs import FULL en multi instance

3.4.1.1 Configuration de la plateforme

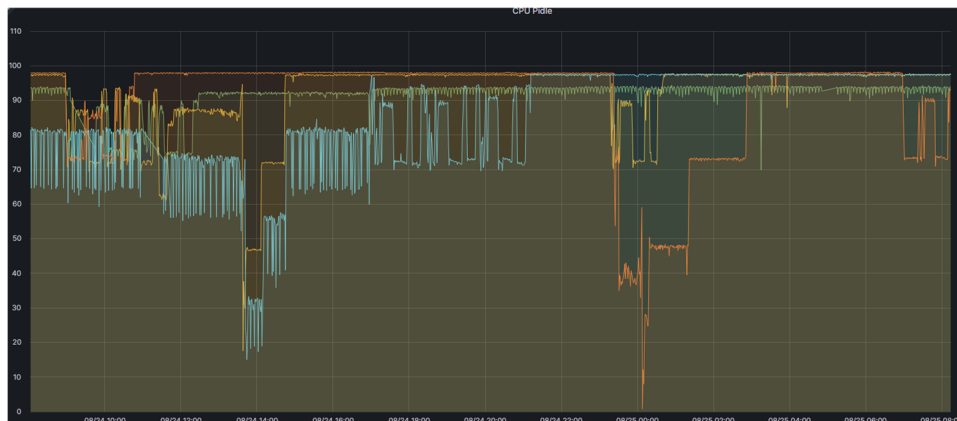
Configuration	Valeur
Nombre de services collecte des données ENT déployés	4 services
Taille JVM	6 144 Mo
Nombre de services import des données ENT déployés	4 services
Taille JVM	10 240 Mo

3.4.1.2 Utilisation CPU machine Applicative

Collecte :

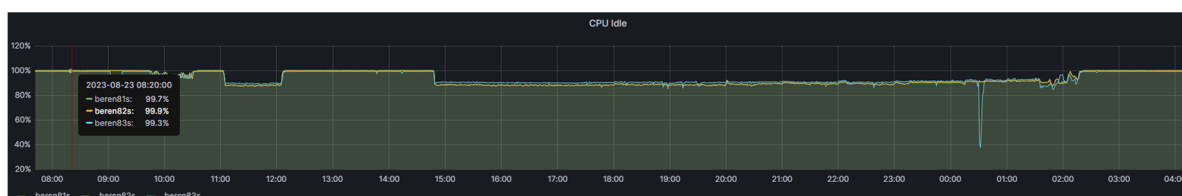


Import :

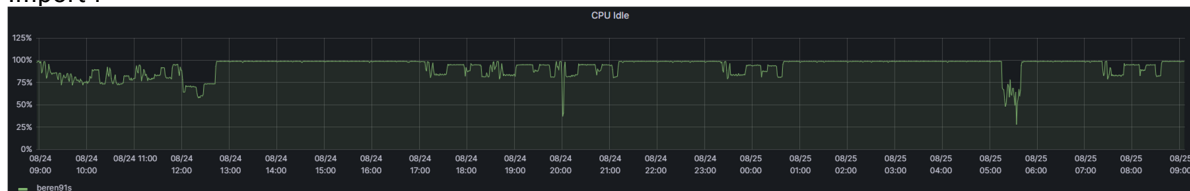


3.4.1.3 Utilisation CPU machine BDD

Collecte :



Import :



3.4.1.4 Résultat du traitement

Configuration	Valeur
Nombre d'archives traités	30
Nombres d'éléments à traiter	182 729 585 éléments
Temps total du traitement de la collecte	13 heures 53 minutes
Nombre d'éléments traités par seconde pour la collecte	3656 éléments/seconde pour la plateforme
Nombres d'éléments à traiter	182 729 585 éléments
Temps total du traitement de la brique d'import	9 heures 15 minutes
Nombre d'éléments traités par seconde pour la brique d'import	5482 éléments/seconde pour la plateforme
Utilisation CPU machine applicative	37 % d'utilisation du CPU en moyenne 80 % d'utilisation du CPU maximum
Utilisation CPU machine BDD	8 % d'utilisation du CPU en moyenne 50 % d'utilisation du CPU maximum

3.4.2 Conclusion import FULL multi instance

Les 4 instances permettent de gérer pour la collecte 3656 éléments/seconde et pour l'import 5482 éléments/seconde. ce qui donne le ratio 2 instances en parallèle permettent de traiter 1.90 fois ce que traite une instance.

En comparaison aux résultats des tirs de performance réalisés en 2020, Sur 4 VM, La brique d'import des données ENT permet d'intégrer 2 fois plus de données qu'une seule VM en 2020.

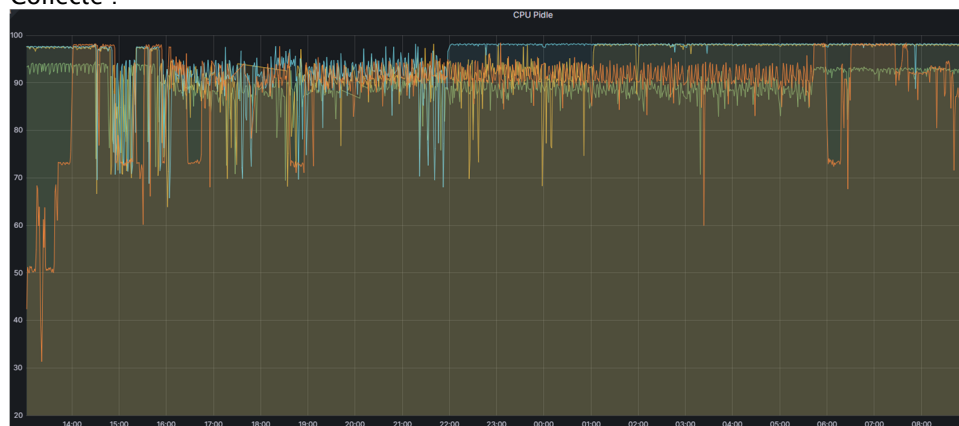
3.4.3 Tirs import DELTA en multiinstance

3.4.3.1 Configuration de la plateforme

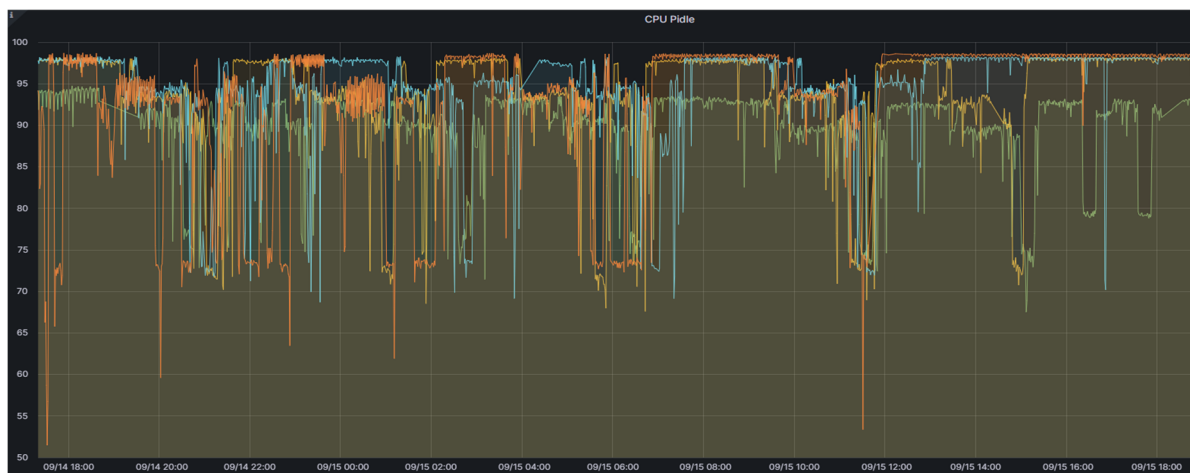
Configuration	Valeur
Nombre de services collecte des données ENT déployés	4 services
Taille JVM	6 144 Mo
Nombre de services import des données ENT déployés	4 services
Taille JVM	10 240 Mo

3.4.3.2 Utilisation CPU machine Applicative

Collecte :

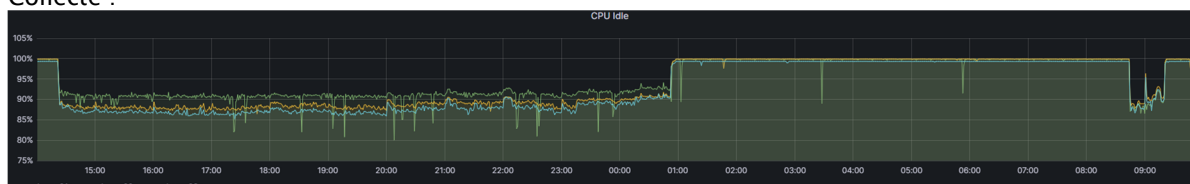


Import :



3.4.3.3 Utilisation CPU machine BDD

Collecte :



Import :



3.4.3.4 Résultat du traitement

Configuration	Valeur
Nombre d'archives traités	30
Nombres d'éléments à traiter	174 875 895 éléments
Temps total du traitement de la collecte	17 heures 42 minutes
Nombre d'éléments traités par seconde pour la collecte	2743 éléments/seconde
Nombres d'éléments à traiter	174 875 895 éléments
Temps total du traitement de la brique d'import	22 heures et 34 minutes
Nombre d'éléments traités par seconde pour la brique d'import	2152 éléments/seconde
Utilisation CPU machine applicative	25 % d'utilisation du CPU en moyenne 60 % d'utilisation du CPU maximum
Utilisation CPU machine BDD	40 % d'utilisation du CPU en moyenne 80 % d'utilisation du CPU maximum

3.4.4 Conclusion import DELTA multi instance

La brique de collecte permet de gérer 2743 éléments/seconde et la brique d'import permet de gérer 2152 éléments/seconde.

En comparaison aux résultats des tirs de performance réalisés en 2020, Sur 4 VM, La brique d'import des données ENT permet d'intégrer environ 2 fois plus de données qu'une seule VM en 2020.

A noter que lors de ce tir, avec 30 archives d'ENT contenant 100% de modification, 3 archives sont tombées en erreur avec des erreurs de « deadlock ». Un nouveau tir sera effectué avec uniquement 15% de modification sur 3 VM.

3.5 Pré-affectation établissement

3.5.1 Description du tir

Le tir utilise le service suivant :

- Le batch de pré-affectation

Le service batch de pré-affectation exécute le job « Pré-affectation s'appuyant sur le type d'affectation établissement ».

Hypothèse du tir :

262 429 accédants sont éligibles à la pré-affectation de ressources pour 10 abonnements différents.

3.5.2 Stratégie du tir

Comme il s'agit d'un batch le tir de montée en charge n'est pas possible.

Un tir est effectué avec la stratégie suivante :

- tir de performance mono instance

3.5.3 Tir mono instance

3.5.3.1 Configuration de la plateforme

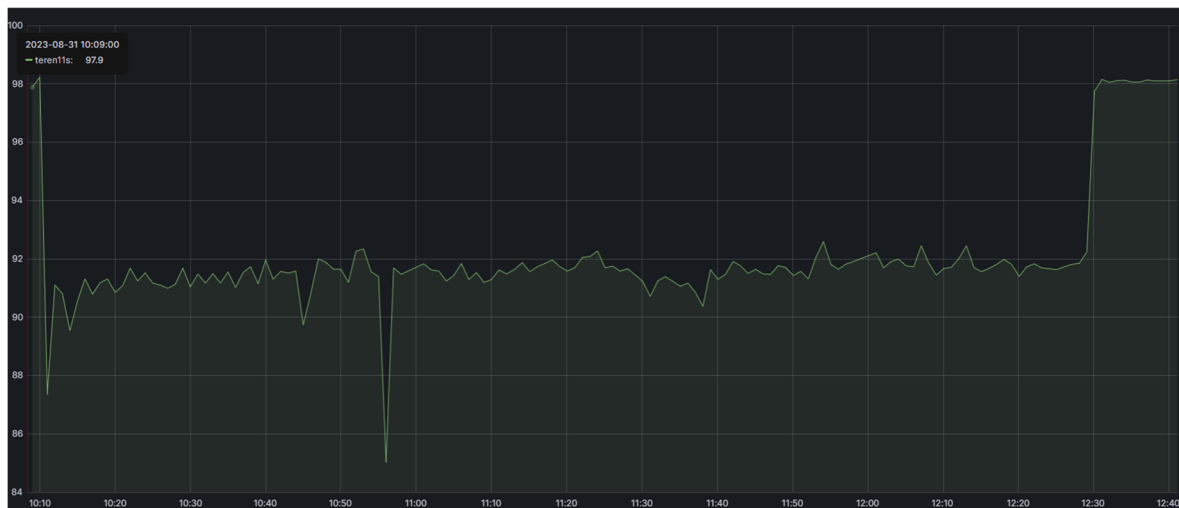
Configuration	Valeur
Nombre de machines déployés	1 machine
Taille JVM	4096 Mo

3.5.3.2 Résultats du tir

Le batch a effectué le traitement avec les métriques suivantes :

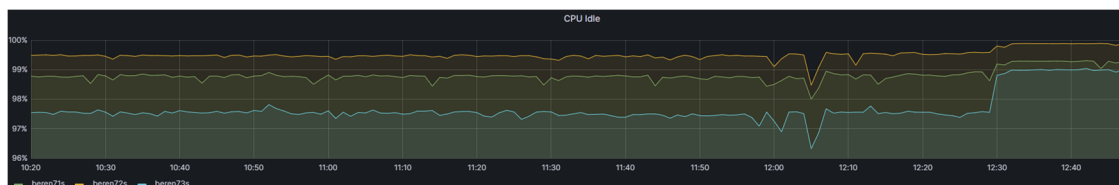
- Création de pré-affectations : 2 624 290
- Temps de traitements : 2h15min

3.5.3.3 Utilisation CPU machine applicative



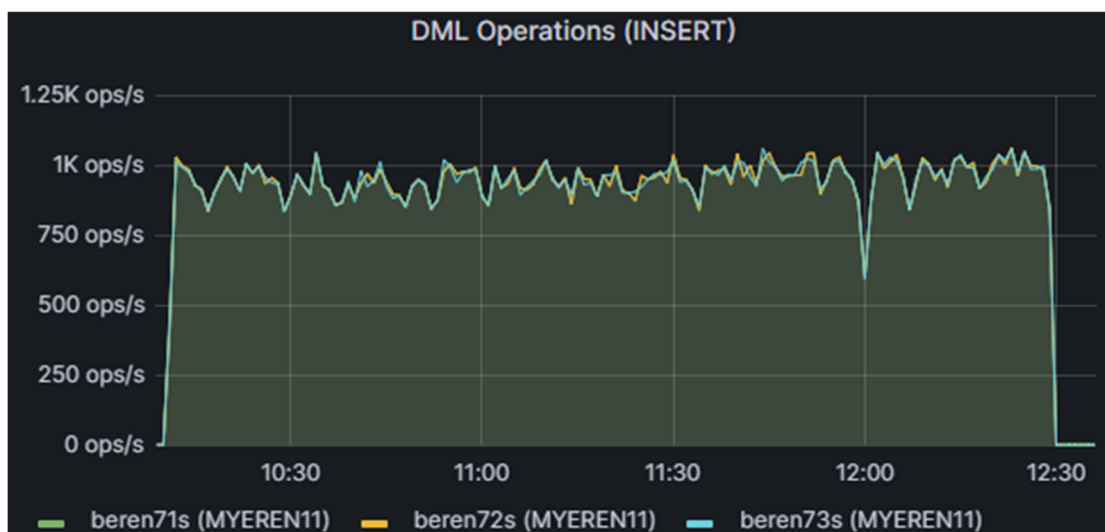
La charge CPU sur le serveur applicatif est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.5.3.4 Utilisation CPU machines BDD



La charge CPU sur les serveurs BDD est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.5.3.5 Requêtes SQL



3.5.4 Conclusion du tir mono instance

Configuration	Valeur
Nombre de création de Pré-affectations	2 624 290
Temps de traitement	2H15
Nombre de pré-affectations par seconde	324 Pré-affectations/s
Gestion d'un pic à 3 millions de pré-affectations	NOK
Utilisation du CPU machine applicative	9 % d'utilisation du CPU en moyenne 17 % d'utilisation du CPU maximum
Utilisation du CPU machines BDD	1 % d'utilisation du CPU en moyenne 2 % d'utilisation du CPU maximum

Ce tir nous permet de vérifier que le service répond aux hypothèses du tir souhaité :

- Gestion d'un pic de 3 millions de pré-affectation

3.6 Pré-affectation niveau éducatif

3.6.1 Description du tir

Le tir utilise le service suivant :

- Le batch de pré-affectation

Le service batch de pré-affectation exécute le job « Pré-Affectation Niveau Educatif pour un nouvel abonnement ».

Hypothèse du tir :

262 429 accédants sont éligibles à la pré-affectation de ressources pour 10 abonnements différents. Le batch est paramétré avec la même configuration que celle de production, c'est-à-dire pré-affectations sur niveau éducatif.

3.6.2 Stratégie du tir

Comme il s'agit d'un batch le tir de montée en charge n'est possible.

Un tir est effectué avec la stratégie suivante :

- tir de performance mono instance

3.6.3 Tirs mono instance

3.6.3.1 Configuration de la plateforme

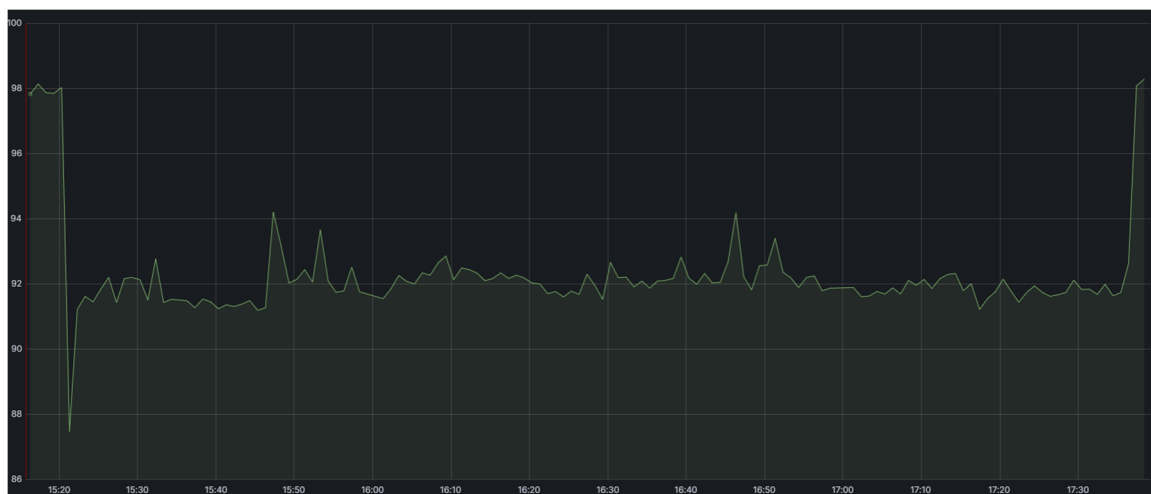
Configuration	Valeur
Nombre de machines déployés	1 machine
Taille JVM	4096 Mo

3.6.3.2 Résultats du tir

Le batch a effectué le traitement avec les métriques suivantes :

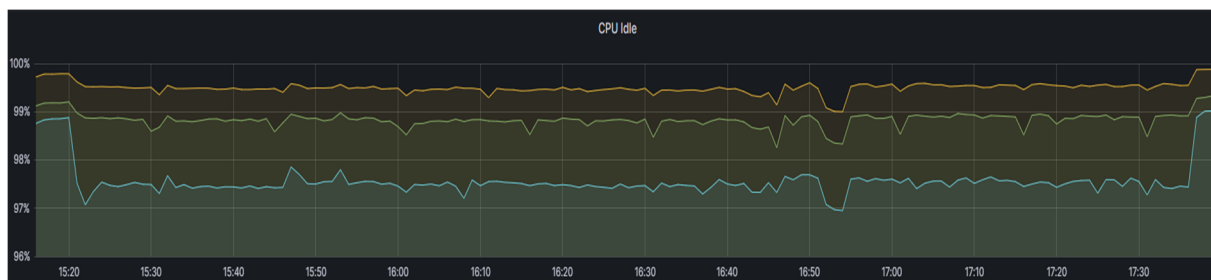
- Pré-affectations : 2 856 867
- Temps de traitements : 2h28min

3.6.3.3 Utilisation CPU machine applicative



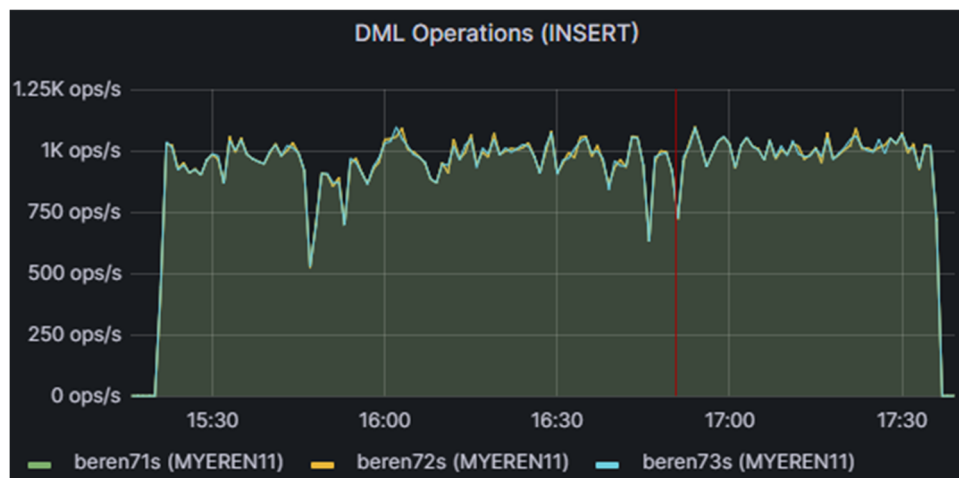
La charge CPU sur le serveur applicatif est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.6.3.4 Utilisation CPU machines BDD



La charge CPU sur les serveurs BDD est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.6.3.5 Requêtes SQL



3.6.4 Conclusion du tir mono instance

Configuration	Valeur
Nombre de création de Pré-affectations	2 856 867
Temps de traitement	2H28
Nombre de pré-affectations par seconde	322 Pré-affectations/s
Gestion d'un pic à 48 M de pré-affectations	OK
Utilisation du CPU machine applicative	8 % d'utilisation du CPU en moyenne 13 % d'utilisation du CPU maximum
Utilisation du CPU machines BDD	1 % d'utilisation du CPU en moyenne 2 % d'utilisation du CPU maximum

3.7 Affectation nouvel arrivant dans un établissement

3.7.1 Description du tir

Le tir utilise le service suivant :

- Le batch de pré-affectation

Le service batch de pré-affectation exécute le job « affectation automatique pour un arrivant dans un établissement ».

Hypothèse du tir :

Le batch va affecter automatiquement 262 429 accédants pour 10 abonnements

3.7.2 Stratégie du tir

Comme il s'agit d'un batch le tir de montée en charge n'est possible.

Un tir est effectué avec la stratégie suivante :

- tir de performance mono instance

3.7.3 Tirs mono instance

3.7.3.1 Configuration de la plateforme

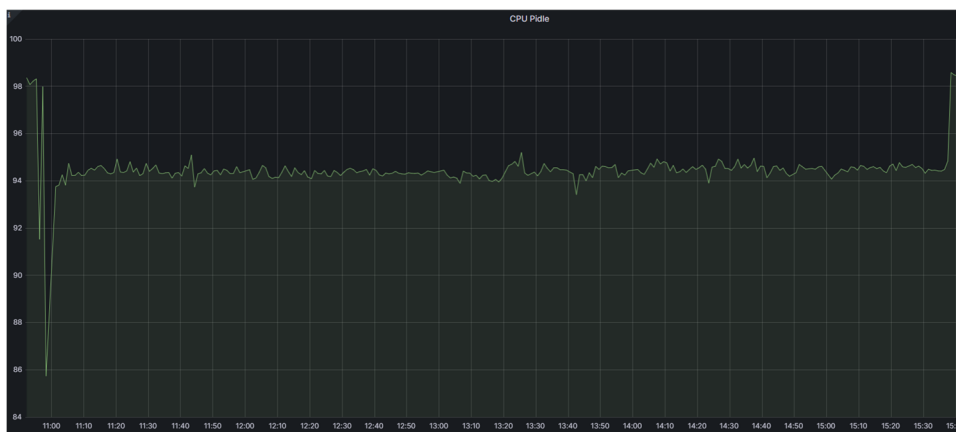
Configuration	Valeur
Nombre de machines déployés	1 machine
Taille JVM	4096 Mo

3.7.3.2 Résultats du tir

Le batch a effectué le traitement avec les métriques suivantes :

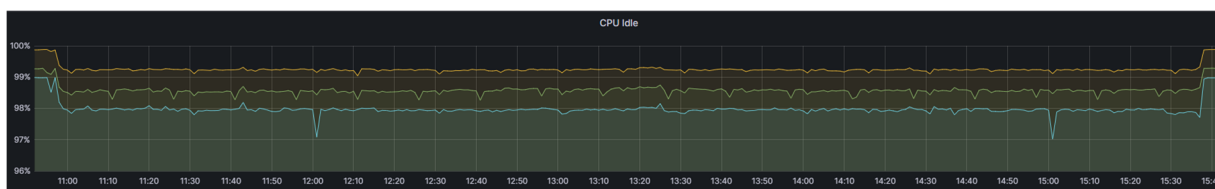
- Affectations automatiques : 439 524
- Temps de traitements : 4h41min

3.7.3.3 Utilisation CPU machine applicative



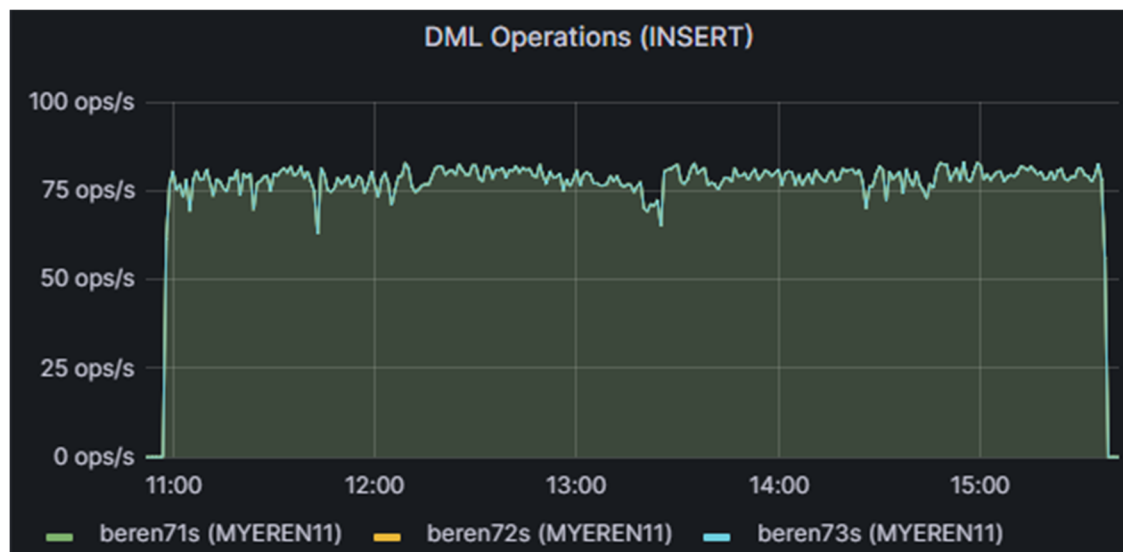
La charge CPU sur le serveur applicatif est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.7.3.4 Utilisation CPU machines BDD



La charge CPU sur les serveurs BDD est en dessous des seuils d'alerte (90% d'utilisation du CPU par exemple) lors de ce tir.

3.7.3.5 Requêtes SQL



3.7.4 Conclusion du tir mono instance

Configuration	Valeur
Nombre de création d'affectations automatiques	439 524
Temps de traitement	4H41
Nombre de pré-affectations par seconde	26 Pré-affectations/s
Gestion d'un pic à 21 000 affectations automatiques	NOK
Utilisation du CPU machine applicative	14 % d'utilisation du CPU maximum
Utilisation du CPU machines BDD	2 % d'utilisation du CPU maximum