

Kierunek: **Informatyka Stosowana (IST)**
Specjalność: **Projektowanie systemów informatycznych (PSI)**

PRACA DYPLOMOWA
MAGISTERSKA

Metody optymalizacji wydajności rozwiązań
ekosystemu Java w ramach AWS Lambda

Piotr Puchała

Opiekun pracy
dr inż. Michał Szczepanik

Słowa kluczowe: AWS Lambda, optymalizacja wydajności, Java, Kotlin, Kotlin
Multiplatform, GraalVM

Streszczenie

Dodaj streszczenie pracy w języku polskim. Staraj się uwzględnić wymienione na stronie tytułowej słowa kluczowe. Uwaga przedstawiony rekomendowany szablon dotyczy pracy dyplomowej pisanej w języku angielskim. W przeciwnym wypadku, student powinien samodzielnie zmienić nazwy „Chapter” na „Rozdział” itp stosując odpowiednie pakiety systemu L^AT_EX oraz ustawienia w pliku *latex-settings.tex*.

Abstract

Streszczenie w języku angielskim.

Spis treści

Wstęp	1
Słownik pojęć i akronimów	3
1 Opis działania modelu FaaS	5
1.1 Model serverless	5
2 Przegląd literatury	7
2.1 Wyniki przeglądu	7
2.1.1 Czynniki wpływające na wydajność funkcji	7

Wstęp

TODO: Do napisania na końcu

Problem badawczy

Usługa AWS Lambda jest jednym z kluczowych serwisów oferowanych przez chmurę Amazon Web Services (AWS) w architekturze bezserwerowej. Wraz z rosnącą popularnością tego rodzaju architektur, pojawia się potrzeba ciągłej poprawy ich działania. Jednym z pierwotnie dostępnych języków programowania w AWS Lambda jest Java (oraz inne języki oparte o Java Virtual Machine), która jednak ze względu na swoją specyfikę (w tym wpływ na responsywność funkcji) nie jest najczęściej wybieraną opcją implementacji w tym serwisie.

Ważnym elementem pracy z AWS Lambda jest wydajność tworzonych funkcji, która przekłada się bezpośrednio na koszt usługi. Wpływ na wydajność mają takie czynniki jak na przykład tzw. zimne i ciepłe starty, czy czas działania samej funkcji, a niewystarczająco szybkie mogą powodować trudności dla programistów.

Mimo rozwoju wielu różnych technologii, ekosystem Java dalej cieszy się dużą popularnością wśród zespołów programistycznych. Poprawa wydajności funkcji AWS Lambda w tym ekosystemie ułatwi programistom decyzję o wyborze tego rozwiązania oraz pozwoli na lepszą pracę z już znanym językiem. Z tego powodu istnieje potrzeba analizy i zaproponowania metod optymalizacji wydajności dla funkcji AWS Lambda w ekosystemie Java.

Cel pracy

Celem pracy jest zaproponowanie nowych metod poprawy wydajności funkcji AWS Lambda w ekosystemie Java oraz analiza ich wpływu na czas działania funkcji i inne wybrane czynniki, które mogą wpłynąć na jakość pracy programistów.

Pytania badawcze

- PB1: Które metody optymalizacji pozwalają na najlepszą poprawę czasu wykonania funkcji AWS Lambda w ekosystemie Java?
- PB2: W jakim stopniu wybrane metody optymalizacji redukują czas zimnego startu funkcji Java w AWS Lambda?
- PB3: Jakie kompromisy w procesie rozwoju oprogramowania wiążą się implementacją poszczególnych metod optymalizacji wydajności funkcji Java w AWS Lambda?

Zakres pracy

Cel pracy zostanie zrealizowany poprzez następujące działania stanowiące zasadniczy wkład pracy:

1. Systematyczny przegląd literatury
2. Identyfikacja i zaproponowanie metod
3. Analiza wpływu

Struktura pracy

TODO: Do napisania na końcu

Słownik pojęć

Function as a Service (FaaS) -
Serverless -
AWS -
AWS Lambda -

1. Opis działania modelu FaaS

1.1. Model serverless

Jednym z dynamicznie rozwijających się obszarów chmur obliczeniowych są usługi serverless. W 2025 roku rynek usług opartych o architekturę bezserwerową jest wart 17,88 miliarda dolarów amerykańskich, a według prognoz jego wartość wzrośnie do 41,14 miliarda w roku 2029 [13]. Badania wykonane na otwartoźródłowych projektach serverless wykazały, że architektura ta używana jest ze względu na niższe koszty, uproszczenie procesów operacyjnych (jak wdrażanie, skalowanie i monitorowanie) oraz bardzo wysoką skalowalność [5]. Cechy te są osiągalne ze względu na wyjątkowe założenia tego modelu.

Przetwarzanie bezserwerowe możemy zdefiniować jako „formę przetwarzania w chmurze, która umożliwia użytkownikom uruchamianie aplikacji sterowanych zdarzeniami i rozliczanych granularnie bez konieczności zarządzania logiką operacyjną” [19]. W definicji tej znajdują się dwa ważne aspekty działania modelu bezserwerowego:

1. „Przetwarzania w chmurze, która umożliwia użytkownikom uruchamianie aplikacji (...) bez konieczności zarządzania logiką operacyjną.”
2. „Aplikacji sterowanych zdarzeniami i rozliczanych granularnie.”

Pierwszy punkt odnosi się do zwiększenia zakresu odpowiedzialności dostawcy chmurowego w porównaniu do klasycznych usług (np. Amazon Elastic Compute Cloud). W usługach tych fizyczne serwery są utrzymywane przez dostawcę chmurowego, a użytkownik jedynie wynajmuje jednostki obliczeniowe. Posiada on dalej kontrolę nad konfiguracją wielu aspektów infrastruktury, co pozwala na większą wydajność rozwoju oprogramowania w porównaniu z środowiskami niechmurowymi. Mimo to, dalej wymaga to poświęcenia czasu i środków na skonfigurowanie oraz zabezpieczenie aplikacji. Architektury bezserwerowe mają na celu uproszczenie tych procesów. Podczas tworzenia aplikacji w usługach bezserwerowych zespoły programistyczne nie muszą zarządzać wdrożeniem, a następnie utrzymaniem serwerów (nawet w formie jednostek jak AWS EC2). Rolą twórcy oprogramowania jest dostarczenie kodu aplikacji lub obrazu Docker [16] [15], które zostaną uruchomione w utrzymywanym przez dostawcę chmurowego środowisku. Dzięki temu inżynierowie mogą skupić się w większym stopniu na logice aplikacji. Pozwala to na zmniejszenie liczby obowiązków, a co za tym idzie kosztów zespołu [20].

Drugi punkt skupia się na charakterystycznym modelu płatności oraz sposobie działania architektur bezserwerowych, który umożliwia taki rodzaj rozliczeń. W przypadku klasycznych usług jak AWS EC2 płatność dokonywana jest za czas działania instancji, niezależnie od tego czy jest ona używana [14]. Model ten

2. Przegląd literatury

2.1. Wyniki przeglądu

W ramach przeglądu wybrano X prac badawczych. Na bazie prac rozpatrzono postawione pytania badawcze. Odpowiedzi na nie zostały zawarte w kolejnych podrozdziałach.

2.1.1. Czynniki wpływające na wydajność funkcji

Pierwszym pytaniem badawczym postawionym do przeglądu literatury jest: „Jakie są główne czynniki wpływające na wydajność funkcji AWS Lambda?”. Identyfikacja czynników wpływających na wydajność jest kluczowa w kontekście jej optymalizacji. Pozwoli to następnie na zrozumienie na które z czynników ma także wpływ twórca funkcji AWS Lambda, co ułatwi dalszą analizę metod poprawy ich wydajności.

Wielkość pamięci funkcji

Kelly, Glavin i Barrett [8] zauważają, że wielkość pamięci funkcji oprócz bezpośredniego wpływu na całkowity czas działania, wywiera także wpływ na inne czynniki jak użycie procesora czy wydajność I/O dysku. W ramach badania wykonano pomiary dla funkcji bezserwerowych oferowanych przez wielu dostawców chmurowych, w tym Amazon Web Services, w celu zrozumienia infrastruktury i jej zarządzania, co domyślnie jest ukryte dla użytkownika. Poprzez analizę maszyn wirtualnych, w ramach których uruchamiany jest kod funkcji, możliwe było otrzymanie wartości parametrów, które nie są domyślnie konfigurowalne podczas wdrażania funkcji przez programistę.

Pomiary pokrywały wiele parametrów funkcji, m. in. łączny czas wykonania, czas inicjalizacji, użycie procesora, wydajność I/O dysku oraz liczbę utworzonych maszyn wirtualnych (co wpływa na częstość zimnych startów). W badaniu uwzględniono predefiniowane wielkości pamięci, które mogą być wybrane przez programistę (128MB, 256MB, 512MB, 1024MB oraz 2048MB). Wraz z wzrostem pamięci funkcji, parametry te poprawiały się.

Autorzy podkreślili ważność odpowiedniego doboru wielkości pamięci podczas tworzenia funkcji. Dodatkowo, wykazali, że w przypadku kolejnych wywołań funkcji, platforma AWS ogranicza ponowne użycie wykorzystanych wcześniej maszyn wirtualnych, co powoduje częstsze zimne starty. Wykazano zatem, że zarówno wielkość pamięci, jak i zimne starty znacząco wpływają na ogólną wydajność funkcji AWS Lambda.

Innym aspektem optymalizacji pamięci jest odpowiednie jej wykorzystanie. W nowoczesnych językach programowania, takich jak Java, powszechne jest użycie różnych implementacji odśmiecania pamięci (ang. garbage collection). Quaresma, Fireman i Pereira [12] przeanalizowali wpływ odśmiecania pamięci w środowisku wykonawczym Java i AWS Lambda. Po pierwsze, wykazali oni, że użycie odśmiecania pamięci może negatywnie wpłynąć na wydajność funkcji. Następnie, poprzez użycie techniki „Garbage Collector Control Interceptor”, złagodzili

negatywny wpływ GC (ang. Garbage Collector), co przyspieszyło czas odpowiedzi o około 10% oraz zmniejszyło koszt działania o 7%.

Wybór odpowiedniej wielkości pamięci funkcji jest bardzo ważnym elementem wdrożenia także ze względu na bezpośredni jej wpływ na koszty. Elgamal i inni autorzy [6] zaproponowali model optymalizacji kosztów funkcji, w którym jednym z czynników była pamięć AWS Lambda. Zwrócili uwagę na to, że nawet niewielkie wartości (z 128 MB do 256 MB) było wstanie poprawić szybkość wykonania algorytmu o 10%, przy jednoczesnym obniżeniu kosztów o 6%.

Pawlik, Figiela i Malawski [11] zwrócili uwagę na wpływ pamięci FaaS dla konkretnych zastosowań naukowych. Dokonali równoczesnej ewaluacji 5120 zadań z użyciem serwisów różnych dostawców (m. in. AWS). Duża liczba zadań wynikała z chęci przetestowania przekroczenia limitów współbieżności oferowanych przez dostawców. Wykazali, że wraz z wzrostem pamięci rośnie wydajność funkcji, mierzona za pomocą wskaźnika GFlops (ang. Giga Floating Point Operations Per Second). Interesującym szczegółem jest niewielka różnica wydajności pomiędzy 2048 MB i 3008 MB (około 0.8%). Autorzy wskazują, że może to być spowodowane taką samą konfiguracją limitów procesora, gdyż wielkość 2048 do niedawna była największą oferowaną przez AWS.

Architektura procesora

Amazon Web Services oferuje możliwość wyboru architektury procesora spośród X86_64 oraz ARM64. Architektura ARM64 jest wspierana poprzez procesory AWS Graviton2 rozwijane przez AWS.

Chen, Hung, Cordingly oraz Lloyd [1] zwrócili uwagę na znaczące różnice wydajności między obiema dostępnymi architekturami procesora. Autorzy przeprowadzili testy wydajnościowe 18 funkcji AWS Lambda i działających na obu rodzajach procesorów (Intel Xeon dla X86_64 oraz AWS Graviton2 dla ARM64).

Wykazali oni podobne zużycie procesora dla obu architektur. Wiele funkcji wykorzystywało wyłącznie jeden z dostępnych rdzeni procesora, co wskazuje na możliwość optymalizacji w kierunku zrównoleglania obliczeń. Mimo podobnego zużycia, sam czas działania funkcji był zróżnicowany. 7 z 18 funkcji działało szybciej na ARM64 (4 były ponad 10% szybsze), podczas gdy 6 działało znacznie wolniej (o ponad 10%). Funkcje ARM64 były bardziej opłacalne dla większości przypadków. 15 z 18 funkcji miało niższe szacunkowe koszty działania na ARM64 w porównaniu do X86 (jednak znaczący wpływ na to miała zniżka oferowana przez dostawcę chmurowego).

Lambion i inni autorzy [9] przeprowadzili analizę użycia algorytmów przetwarzania języka naturalnego z użyciem obu architektur procesora. W ramach badania przygotowali oni składający się z kilku etapów pipeline, który używał wspomnianych algorytmów. Funkcje zostały wdrożone z użyciem obu architektur oraz w różnych regionach AWS.

Wydajność obu architektur różniła się w zależności od etapu pipeline'u. Dla regionu us-east-2 ARM64 był szybszy w przypadku funkcji przetwarzania wstępnego (o 7,3%) i zapytań (o 8,9%), podczas gdy X86_64 był znacznie szybszy (o 23,6%) w przypadku funkcji treningowej. W ujęciu globalnym funkcje ARM64 były średnio o 1.7% szybsze niż funkcje X86_64.

Działanie funkcji różniło się także w zależności od regionu. Funkcje w architekturze X86_64 działały najszybciej w regionie eu-central-1, a najwolniej w us-west-2. W przypadku ARM64, region us-west-2 był najszybszy, a us-east-2 najwolniejszy. Funkcje wykazały także tendencję

do bycia szybszymi poza typowymi godzinami pracy (np. funkcje w godzinach 6:00-8:00 działały o 6% szybciej niż funkcje w godzinach 10:00-12:00).

Zimne starty

TODO: [10] TODO: [18] TODO: [4] TODO: [3]

Język programowania

TODO: [7] TODO: [17] TODO: [2]

Bibliografia

- [1] X. Chen, L.-H. Hung, R. Cordingly, and W. Lloyd. X86 vs. arm64: An investigation of factors influencing serverless performance. In *Proceedings of the 9th International Workshop on Serverless Computing*, WoSC '23, page 7–12, New York, NY, USA, 2023. Association for Computing Machinery.
- [2] R. Cordingly, H. Yu, V. Hoang, D. Perez, D. Foster, Z. Sadeghi, R. Hatchett, and W. J. Lloyd. Implications of programming language selection for serverless data processing pipelines. page 704 – 711, 2020. Cited by: 21.
- [3] J. Dantas, H. Khazaei, and M. Litoiu. Application deployment strategies for reducing the cold start delay of aws lambda. In *2022 IEEE 15th International Conference on Cloud Computing (CLOUD)*, pages 1–10, 2022.
- [4] A. Ebrahimi, M. Ghobaei-Arani, and H. Saboohi. Cold start latency mitigation mechanisms in serverless computing: Taxonomy, review, and future directions. *Journal of Systems Architecture*, 151:103115, 2024.
- [5] S. Eismann, J. Scheuner, E. van Eyk, M. Schwinger, J. Grohmann, N. Herbst, C. L. Abad, and A. Iosup. Serverless applications: Why, when, and how? *IEEE Software*, 38(1):32–39, 2021.
- [6] T. Elgamal, A. Sandur, K. Nahrstedt, and G. Agha. Costless: Optimizing cost of serverless computing through function fusion and placement. In *2018 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 300–312, 2018.
- [7] D. Jackson and G. Clynych. An investigation of the impact of language runtime on the performance and cost of serverless functions. In *2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion)*, pages 154–160, 2018.
- [8] D. Kelly, F. Glavin, and E. Barrett. Serverless computing: Behind the scenes of major platforms. In *2020 IEEE 13th International Conference on Cloud Computing (CLOUD)*, pages 304–312, 2020.
- [9] D. Lambion, R. Schmitz, R. Cordingly, N. Heydari, and W. Lloyd. Characterizing x86 and arm serverless performance variation: A natural language processing case study. In *Companion of the 2022 ACM/SPEC International Conference on Performance Engineering*, ICPE '22, page 69–75, New York, NY, USA, 2022. Association for Computing Machinery.
- [10] J. Manner, M. Endreß, T. Heckel, and G. Wirtz. Cold start influencing factors in function as a service. In *2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion)*, pages 181–188, 2018.

- [11] M. Pawlik, K. Figiela, and M. Malawski. Performance considerations on execution of large scale workflow applications on cloud functions, 2019.
- [12] D. Quaresma, D. Fireman, and T. E. Pereira. Controlling garbage collection and request admission to improve performance of faas applications. In *2020 IEEE 32nd International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, pages 175–182, 2020.
- [13] Research and Markets. Serverless architecture market report 2025, (2025). Dostęp 20.03.2025 z <https://www.researchandmarkets.com/reports/5953253/serverless-architecture-market-report>.
- [14] A. W. Services. Amazon ec2 documentation, 2025. Dostęp 21.03.2025 z <https://docs.aws.amazon.com/ec2/>.
- [15] A. W. Services. Amazon ecs developer guide, 2025. Dostęp 21.03.2025 z <https://docs.aws.amazon.com/AmazonECS/latest/developerguide>.
- [16] A. W. Services. Aws lambda documentation, 2025. Dostęp 21.03.2025 z <https://docs.aws.amazon.com/lambda/latest/dg/>.
- [17] S. Shrestha. Comparing programming languages used in aws lambda for serverless architecture. Bachelor’s thesis, Theseus, 2019.
- [18] P. Vahidinia, B. Farahani, and F. S. Aliee. Cold start in serverless computing: Current trends and mitigation strategies. In *2020 International Conference on Omni-layer Intelligent Systems (COINS)*, pages 1–7, 2020.
- [19] E. van Eyk, A. Iosup, C. L. Abad, J. Grohmann, and S. Eismann. A spec rg cloud group’s vision on the performance challenges of faas cloud architectures. In *Companion of the 2018 ACM/SPEC International Conference on Performance Engineering*, page 21–24, New York, NY, USA, 2018. Association for Computing Machinery.
- [20] J. Wen, Z. Chen, X. Jin, and X. Liu. Rise of the planet of serverless computing: A systematic review. *ACM Trans. Softw. Eng. Methodol.*, 32(5), 2023.