MINISTERUL EDUCAŢIEI ȘI CERCETĂRII
AL REPUBLICII MOLDOVA
Universitatea Tehnică a Moldovei
Facultatea Electronică şi Telecomunicaţii
Departamentul Telecomunicaţii Și Sisteme Electronice

Admis la susţinere

Şefă departament:
Tîrșu Valentina conf. univ., dr.

_____

„____ ” _____2024

PROIECTAREA SERVERULUI CLOUD MULTIFUCȚIONAL DE ÎNALTĂ PERFORMANȚĂ CU ACCES PRIVAT

Teză de licenţă

Student: **Fachira Nicolae IMTC-201**

Coordonator: **Cristea Ecaterina asist. univ.**

Consultant: **Grițco Maria asist. univ.**

Chişinău, 2024

ADNOTARE

Fachira Nicolae, studentul grupei IMTC-201

Tema: Proiectarea serverului cloud multifunctional de înaltă performanță cu acces privat.

Cuvinte cheie: GPU, cluster, Chroma Vector DB, NVSwitch, Kubernetes, Docker, AI

Scopul lucrării: Proiectarea și implementarea unui cluster GPU capabil să ofere performanțe înalte pentru sarcini de calcul intensive, utilizând Chroma Vector DB pentru crearea embeddingurilor.

Obiectivele lucrării:

- Proiectarea infrastructurii hardware a clusterului GPU
- Configurarea și optimizarea rețelei de interconectare a serverelor
- Implementarea și configurarea Chroma Vector DB
- Automatizarea gestionării resurselor utilizând Kubernetes și Docker
- Evaluarea economică a investiției și analizarea beneficiilor pe termen lung.

Metodele aplicate la elaborarea lucrării: Utilizarea tehnologiilor de top în calculul distribuit, precum Kubernetes și Docker pentru orchestrarea containerelor, NVSwitch pentru interconectarea GPU-urilor și Chroma Vector DB pentru gestionarea bazelor de date vectoriale.

Rezultatele obținute: Drept urmare a implementării clusterului GPU, s-a demonstrat o creștere semnificativă a performanței în sarcinile de calcul intensive. Arhitectura propusă a permis o scalabilitate și flexibilitate sporită, asigurând în același timp o gestionare eficientă a resurselor și o securitate robustă. Analiza economică a relevat o perioadă de recuperare a investiției de aproximativ 6.88 ani, confirmând viabilitatea financiară a proiectului. Aceste rezultate subliniază importanța utilizării tehnologiilor moderne pentru optimizarea infrastructurii de calcul și pentru maximizarea randamentului investițiilor în domeniul IT.

ANNOTATION

Author: Fachira Nicolae. IMTC group – 201

Topic: Design of a High-Performance Multifunctional Private Cloud Server

Keywords: GPU, cluster, Chroma Vector DB, NVSwitch, Kubernetes, Docker, AI.

Purpose of the Work: The design and implementation of a high-performance GPU cluster capable of providing intensive computational performance, using Chroma Vector DB for creating embeddings.

Objectives of the Work:

- Design of the hardware infrastructure of the GPU cluster
- Configuration and optimization of the network interconnection of servers
- Implementation and configuration of Chroma Vector DB
- Automation of resource management using Kubernetes and Docker
- Economic evaluation of the investment and analysis of long-term benefits.
-

Methods Applied in the Work: Utilization of top-tier technologies in distributed computing, such as Kubernetes and Docker for container orchestration, NVSwitch for GPU interconnection, and Chroma Vector DB for managing vector databases.

Results Obtained: As a result of the GPU cluster implementation, a significant increase in performance for intensive computational tasks was demonstrated. The proposed architecture allowed for enhanced scalability and flexibility, ensuring efficient resource management and robust security. The economic analysis revealed a payback period of approximately 6.88 years, confirming the financial viability of the project. These results highlight the importance of using modern technologies to optimize computing infrastructure and maximize the return on IT investments.

CUPRINS

INTRODUCERE……..……4

1 NOȚTIUNI GENERALE DESPRE INFRASTRUCTURA DE SERVERE, DESIGN DE SISTEM ȘI TEHNOLOGIILE UTILIZATE.....5

1.1 Designul Sistemelor.....5

1.1.1 Conceptul de design de sistem 5

1.1.2 Arhitectura sistemelor distribuite.....9

1.2 Termeni și Tehnologii Utilizate.....21

1.2.1 GPU Computing.....21

1.2.2 Clustere de Servere.....22

1.2.3 Containere și Orchestrare.....24

1.2.4 Baze de Date Vectoriale.....30

1.2.5 Load Balancing.....32

1.2.6 Monitorizare și Logging.....33

1.2.7 Securitate.....35

2 PROIECTAREA SISTEMULUI GPU COMPUTING CLUSTER IN ACORD CU PARADIGMELE CLOUD NATIVE.....38

2.1 Planificarea și Configurarea Hardware.....38

2.1.1 Descrierea infrastructurii hardware.....38

2.1.2 Instalarea și configurarea hardware-ului.....48

2.2 Configurarea Sistemului de Operare și configurarea echipamentului de networking,.....52

2.2.1 Instalarea Ubuntu Server.....52

2.2.2 Configurarea rețelei pe servere.....53

2.2.3 Configurarea rețelei pe echipamentul de retea.....55
2.2.4 Optimizarea sistemului de operare pentru performanță GPU.....58
2.3 Implementarea Software-ului.....60
2.3.1 Instalarea și configurarea Chroma Vector DB.....61
2.3.2 Configurarea API-ului și a accesului SSH.....62
2.3.3 Managementul utilizatorilor și al sesiunilor.....64
2.4 Gestionarea Resurselor și Load Balancing.....66
2.4.1 Configurarea load balancer-ului HAProxy.....66
2.4.2 Gestionarea Resurselor Dinamice (Kubernetes și Docker).....68
2.4.3 Monitorizarea și Ajustarea Încărcării între Servere(Prometheus, Grafana).....70
2.4.4 Arhitectura Software și Flow-ul de Date.....72
2.5 Securitatea și sănătatea în muncă.....75
3. IDENTIFICAREA EFICIENȚEI ECONOMICE A GPU CLUSTER-ULUI
3.1 Calculul cheltuerilor de achizitionare și intreținere.....77
3.2 Determinarea volumului de investiții pentru fiabilitate.....78
3.3 Termenul de recuperare a investiției.....80
CONCLUZIE.....81
BIBILIOGRAFIE.....82

INTRODUCERE

În era digitală contemporană, creșterea exponențială a volumului de date și complexitatea aplicațiilor informatice au impus dezvoltarea unor soluții tehnologice avansate pentru procesarea eficientă și rapidă a informațiilor. În acest context, sistemele de calcul bazate pe unități de procesare grafică (GPU) au devenit esențiale datorită capacității lor de a efectua calcule paralele masive. Aceste sisteme sunt utilizate într-o varietate de domenii, de la inteligență artificială și învățare automată, până la simulări științifice și prelucrarea datelor mari.

Lucrarea de față își propune să proiecteze un cluster de servere GPU cu acces global, destinat utilizatorilor din întreaga lume, având ca principal serviciu crearea de embedding-uri pentru Chroma Vector DB. Acest serviciu va fi accesibil atât prin intermediul unui API, cât și prin conexiuni SSH la terminalul serverului. Proiectul va include patru servere GPU montate în rack-uri 8U, echipate cu SYS-821GE-TNHR și fiecare dotat cu HGX H100 8-GPU, precum și un rack mount 1U dedicat gestionării sesiunilor, stocării informațiilor utilizatorilor și echilibrării încărcării între servere.

Pentru a asigura o funcționare optimă și scalabilă a sistemului, vom utiliza o varietate de tehnologii și practici moderne. Sistemul de operare ales este Ubuntu Server, datorită stabilității și flexibilității sale. Stocarea vectorilor se va realiza utilizând ChromaDB, o soluție eficientă pentru baze de date vectoriale. De asemenea, vom implementa containere și orchestrare cu Docker și Kubernetes, pentru a facilita izolarea și gestionarea serviciilor. Pentru echilibrarea încărcării vom folosi soluții precum HAProxy sau NGINX, iar monitorizarea și logging-ul vor fi asigurate de Prometheus, Grafana și ELK Stack.

Securitatea este un aspect esențial al acestui proiect, de aceea vom implementa practici și tehnologii adecvate, precum OpenSSH pentru accesul la terminal, Keycloak pentru gestionarea autentificării și autorizării și UFW pentru configurarea firewall-ului.

Această lucrare va aborda, în primul rând, fundamentarea teoretică necesară pentru înțelegerea și proiectarea unui astfel de sistem complex. Vom explora conceptele de design de sistem, arhitectura sistemelor distribuite și tehnologiile specifice utilizate. În continuare, vom detalia aspectele practice ale proiectării și implementării clusterului GPU, incluzând topologia fizică și logica sistemului, organizarea software-ului, gestionarea instanțelor și utilizatorilor, monitorizarea și măsurile de securitate. Scopul principal al acestei lucrări este de a demonstra cum pot fi integrate și utilizate tehnologiile moderne pentru a crea un sistem de calcul performant, scalabil și securizat, capabil să răspundă cerințelor utilizatorilor globali. Lucrarea se încheie cu o discuție asupra rezultatelor obținute și a posibilităților de dezvoltare ulterioară a sistemului.

BIBLIOGRAFIE

[1] „AWS Database Blog” [Online]. [citat 14.01.2024] Disponibil: https://aws.amazon.com/ru/blogs/database/understand-and-build-a-hybrid-database-with-amazon-rds-and-aws-outposts/

[2] „MongoDB Database Sharding” [Online]. [citat 25.01.2024] Disponibil: https://www.mongodb.com/resources/products/capabilities/database-sharding-explained

[3] „Medium forum, Saurabh Goyal post” [Online]. [citat 26.01.2024] Disponibil: https://medium.com/delta-exchange/centralized-vs-decentralized-vs-distributed-41d92d463868

[4] „Modex: Case Study” [Online]. [citat 29.01.2024] Disponibil: https://www.modex.tech/blog/centralized-vs-decentralized-vs-distributed-systems

[5] „DEV Community, Leonardo Luís Dalcegio post” [Online]. [citat 08.02.2024] Disponibil: https://dev.to/leodalcegio/a-comprehensive-guide-to-mapreduce-distributed-data-processing-3lj8

[6] „Telefonaktiebolaget LM Ericssonm, Julien Forgeat post” [Online]. [citat 11.02.2024] Disponibil: https://www.ericsson.com/en/blog/2015/11/data-processing-architectures--lambda-and-kappa

[7] Damian Wojsław, Grzegorz Adamowicz, 2023. „The Linux DevOps Handbook*”*

[8] William Shotts, 2019. „The Linux Command Line 2ND Edition*”*

[9] „Databricks Inc.” [Online]. [citat 20.02.2024] Disponibil: https://www.databricks.com/glossary/hadoop-distributed-file-system-hdfs

[10] „ Amazon Web Services, Inc.” [Online]. [citat 11.04.2024] Disponibil: https://aws.amazon.com/ru/compare/the-difference-between-rabbitmq-and-kafka/

[11] James Freeman, 2020. „ *Hands-On Enterprise Automation on Linux”*, CRC Press

[12] „ Wikimedia Foundation, Inc.” [Online] [citat 11.04.2024]. Disponibil: https://en.wikipedia.org/wiki/CAP_theorem

[13] Martin Kleppmann, 2015. „ *A Critique of the CAP Theorem”*, University of Cambridge

[14] Mohan Goppalakrishnan, 2024. *„Introduction to GEN2 (BB5216)”,* Ericsson

[15] „ NVIDIA Corporation, William Tsu” [Online] [citat 16.04.2024]. Disponibil: https://developer.nvidia.com/blog/introducing-nvidia-hgx-h100-an-accelerated-server-platform-for-ai-and-high-performance-computing/

[16] NVIDIA Corporation, 2012. , *“ OpenCL Programming for the CUDA Architecture”,* NVIDIA

[17] „ AccuWeb.Cloud” [Online]. Disponibil: https://accuweb.cloud/blog/server-clustering-with-types-and-benefits/

[18] „ YouTube, jscapeus” [Online]. Disponibil:

https://www.youtube.com/watch?app=desktop&v=MDuCFh1XuZU&ab_channel=jscapeus

[19] “Docker Inc.” „*Docker Documentation”* [Online]. Disponibil: https://docs.docker.com/get-started/overview/

[20] „Simform”,” Kubernetes Architecture and Components” [Online]. [citat 29.01.2024] Disponibil: https://www.simform.com/blog/kubernetes-architecture/

[21] „Chroma”,” Chroma Documentation” [Online]. [citat 14.04.2024] Disponibil: https://docs.trychroma.com/

[22] „Medium”, “Unveiling the Architectural Brilliance of Prometheus” [Online]. [citat 13.04.2024] Disponibil: https://medium.com/@extio/unveiling-the-architectural-brilliance-of-prometheus-af07cca14896

[23] „ Red Hat, Inc.”, “Security by design: Security principles and threat modeling” [Online]. [citat 11.03.2024] Disponibil: https://www.redhat.com/en/blog/security-design-security-principles-and-threat-modeling

[24] „ EDB”, “EDB Postgressdistributed” [Online]. [citat 11.03.2024] Disponibil: https://www.enterprisedb.com/docs/pgd/latest/planning/architectures/