Možnosti monitoringu a optimalizace HPC clusterů s GPU kartami

Tomáš Huťa

Bakalářská práce
2022

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky
Ústav bezpečnostního inženýrství

Akademický rok: 2021/2022

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Tomáš Huťa**
Osobní číslo:	**A19263**
Studijní program:	**B3902 Inženýrská informatika**
Studijní obor:	**Bezpečnostní technologie, systémy a management**
Forma studia:	**Kombinovaná**
Téma práce:	**Možnosti monitoringu a optimalizace HPC clusterů s GPU kartami**
Téma práce anglicky:	**Possibilities of Monitoring and Optimization of HPC Clusters with GPU Cards**

Zásady pro vypracování

1. Specifikujte možnosti výběru komponent pro HPC a GPU clustery.
2. Navrhněte možnosti monitoringu těchto clusterů.
3. Proveďte návrh 3 systémů pro využití v AI a Deep Learningu.
4. Porovnejte výkonnost a efektivitu navržených systémů.
5. Zpracujte možnosti optimalizace využití zdrojů Vašich návrhů v oblasti AI a DL.
6. Proveďte zabezpečení systémů před neoprávněným přístupem a útoky ze sítě Internet.

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam doporučené literatury:

1. ZHONG, Li, Dennis HOPPE, Naweiluo ZHOU a Oleksandr SHCHERBAKOV. Hybrid workflow of Simulation and Deep Learning on HPC: A Case Study for Material Behavior Determination. _2021 IEEE International Conference on Cluster Computing (CLUSTER), Cluster Computing (CLUSTER), 2021 IEEE International Conference on, CLUSTER_ [online]. 2021, , 698-704 [cit. 2021-12-01]. ISBN 9781728196664. ISSN 21689253. Dostupné z: doi:10.1109/Cluster48925.2021.00104
2. KOLOUCH, Jan a Pavel BAŠTA. CyberSecurity. Praha: CZ.NIC, 2019. ISBN 978-80-88168-31-7.
3. KOLOUCH, Jan. CyberCrime. Praha: CZ.NIC, 2016. ISBN 978-80-88168-15-7.
4. SELECKÝ, Matúš. Penetrační testy a exploitace. Brno: Computer Press, 2012. ISBN 978-80-251-3752-9.
5. ŠULC, Vladimír. Kybernetická bezpečnost. Plzeň: Aleš Čeněk, 2018. ISBN 978-80-7380-737-5.

Vedoucí bakalářské práce: **Ing. David Malaník, Ph.D.**
Ústav informatiky a umělé inteligence

Datum zadání bakalářské práce: **17. ledna 2022**
Termín odevzdání bakalářské práce: **31. května 2022**

UNIVERZITA TOMÁŠE BATI VE ZLÍNĚ
Fakulta aplikované informatiky
L.S.

doc. Mgr. Milan Adámek, Ph.D. v.r.
děkan

Ing. Jan Valouch, Ph.D. v.r.
ředitel ústavu

Ve Zlíně dne 17. ledna 2022

Prohlašuji, že

- beru na vědomí, že odevzdáním bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby;
- beru na vědomí, že bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk bakalářské práce bude uložen v příruční knihovně Fakulty aplikované informatiky Univerzity Tomáše Bati ve Zlíně;
- byl/a jsem seznámen/a s tím, že na moji bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- beru na vědomí, že podle § 60 odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- beru na vědomí, že podle § 60 odst. 2 a 3 autorského zákona mohu užít své dílo – bakalářskou práci nebo poskytnout licenci k jejímu využití jen připouští-li tak licenční smlouva uzavřená mezi mnou a Univerzitou Tomáše Bati ve Zlíně s tím, že vyrovnání případného přiměřeného příspěvku na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše) bude rovněž předmětem této licenční smlouvy;
- beru na vědomí, že pokud bylo k vypracování bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tedy pouze k nekomerčnímu využití), nelze výsledky bakalářské práce využít ke komerčním účelům;
- beru na vědomí, že pokud je výstupem bakalářské práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá. Neodevzdání této součásti může být důvodem k neobhájení práce.

Prohlašuji,

- že jsem na bakalářské práci pracoval samostatně a použitou literaturu jsem citoval. V případě publikace výsledků budu uveden jako spoluautor.
- že odevzdaná verze bakalářské práce a verze elektronická nahraná do IS/STAG jsou totožné.

Ve Zlíně, dne 24.05.2022

Tomáš Huťa, v.r.
podpis studenta

ABSTRAKT

Cílem mé práce je zmapovat aktuální výpočetní zařízení na trhu, jež je možné využít k provozu systémů zaměřených na AI a Deep learning. Následně navrhnout vlastní konfiguraci výpočetního serveru spolu s možností monitoringu všech typů serverů, ať už vlastní konfigurace či zakoupených profesionálních zařízení. Dále pak systémy určené pro práci s AI a Deep learningem.

Praktická část je zaměřená na porovnání výkonu a spotřeby mezi grafickými kartami a zpracování možností optimalizace spotřeby či zdrojů grafických karet. Rovněž jsou v práci zpracované i metody zabezpečení serveru před neoprávněnými přístupy útočníků či útoky ze sítě Internet.

Klíčová slova: AI, Deep learning, cluster, hardware, grafická karta, monitoring, optimalizace, zabezpečení

ABSTRACT

The aim of my work is to map the current computing devices on the market that can be used to run AI and Deep learning systems. Subsequently, I will propose a custom computing server configuration along with options for monitoring all types of servers, whether self-configured or purchased professional equipment. Furthermore, systems designed to work with AI and Deep learning.

The practical part is focused on comparing performance and power consumption between graphics cards and discussing the possibilities of optimizing power consumption or resources of graphics cards. Also the methods of server security against unauthorized accesses of attackers or attacks from the Internet are elaborated in the thesis.

Keywords: AI, Deep learning, cluster, hardware, graphic card, monitoring, optimization, security

Poděkování

Chtěl bych poděkovat svému vedoucímu Ing. Davidu Malaníkovi, Ph.D. za odborné a cenné rady a velmi profesionální přístup, kterým mě provázel po celou dobu mé práce. Dále bych chtěl poděkovat své rodině za podporu po celou dobu mého studia.

Prohlašuji, že odevzdaná verze bakalářské práce a verze elektronická nahraná do IS/STAG jsou totožné.

OBSAH

ÚVOD

Stále častěji se můžeme potkávat s technologiemi, které nějakým způsobem pracují s umělou inteligencí. Rozumím tím technologie založené na rozpoznávání hlasu, obličeje, kamery schopné zachytit blížící se vozidlo a rozpoznat typ daného vozidla, ale také generátory, které za pomoci umělé inteligence dokážou vytvářet podobizny člověka, zvířat či náhodné obrázky téměř všeho. Všechny tyto technologie by se ovšem neobešly bez kvalitního základu ve výpočetní technice. Systém je potřeba nejprve nějakým způsobem naučit, seznámit s tím, jak dané situace nebo předměty vypadají, než bude schopen pracovat samostatně s daty na základě toho, co se naučí. A k tomu je zapotřebí využít výpočetní výkon.

Teoretická část bakalářské práce se skládá celkem z pěti kapitol a zaměřuje se na možnosti výběru hardwaru pro stavbu velmi výkonných počítačových sestav sloužících k provozu výpočetně náročných systémů určených pro práci s AI a Deep learningem. Dále se práce zaměřuje na možnosti monitoringu těchto výpočetních stanic, zejména na provozní monitoring stavu jednotlivých komponent a jejich využití. Součástí je i návrh systémů implementovatelných na výpočetní stanici a jejich využití v oblasti umělé inteligence a strojového učení.

V praktické části práce se zabývám sestavením výpočetního serveru z vybraných hardwarových komponentů a jejich základního nastavení. Dále bylo nutné na server nainstalovat operační systém, následně provést porovnání výkonnosti a efektivity systémů a provést návrh optimalizace spotřeby elektrické energie grafických karet. V neposlední řadě je potřeba takovou výpočetní stanici ochránit před neoprávněnými přístupy, jak z internetu, tak z lokální sítě. Proto je v práci obsaženo i zabezpečení výpočetního serveru na systémové úrovni.

I. TEORETICKÁ ČÁST

1 HPC CLUSTER

1.1 Opodstatnění tvoření clusterů

Téměř každý běžný počítač je dostatečně výkonný na kancelářskou práci a surfování po internetu. Pokud ale budeme chtít provozovat výpočetně náročné systémy, například na provádění renderingu animovaných scén, strojového učení, simulace počasí či provozu umělé inteligence, bude zapotřebí mít několikanásobně výkonnější počítač, než jaký máme běžně k dispozici. Takový počítač bude nejen velmi drahý, ale také náchylný na poruchy, jelikož se spoléháme na provoz jednoho jediného stroje. V průmyslovém využití by nákup takového počítače nedával smysl. Mnohem lepší myšlenkou je spojit více počítačů do jednoho a za pomocí počítačové sítě tak vytvořit skupinu mnoha počítačů.

1.2 Typy clusterů

Clustery řadíme do několika typů, podle plnění svých funkcí.

1.2.1 Výpočetní cluster (High performance computing)

Spojením několika počítačů do sebe za pomoci počítačové sítě nám vznikne výpočetní uzel. Tímto způsobem dosáhneme nižší pořizovací ceny za hardware, než při použití jednoho supervýkonného počítače

1.2.2 Cluster s vysokou dostupností (High availability cluster)

Spojení několika počítačů do sebe získáváme potřebnou zastupitelnost každého z nich. V případě poruchy jednoho z počítačů jej zastoupí jiný, připravený. Využití najdeme v provozu systémů, kde je zapotřebí vysoká dostupnost (např. databázové systémy, skladové systémy)

1.2.3 Cluster s rozložením zátěže (Load balancing)

Na jednotlivých počítačích jsou provozovány paralelně totožné systémy, jejichž cílem je snížit celkové zatížení systému. Tento typ clusteru se používá například k provozu velkých e-shopů, kde by provoz na jeden počítač byl příliš náročný a nákladný.

1.2.4 Cluster se zaměřením na úložnou kapacitu (Storage cluster)

Cílem toho řešení je snížit zátěž na diskové úložiště, zvýšit tak jeho spolehlivost a zároveň snížit riziko selhání diskového pole. [1]

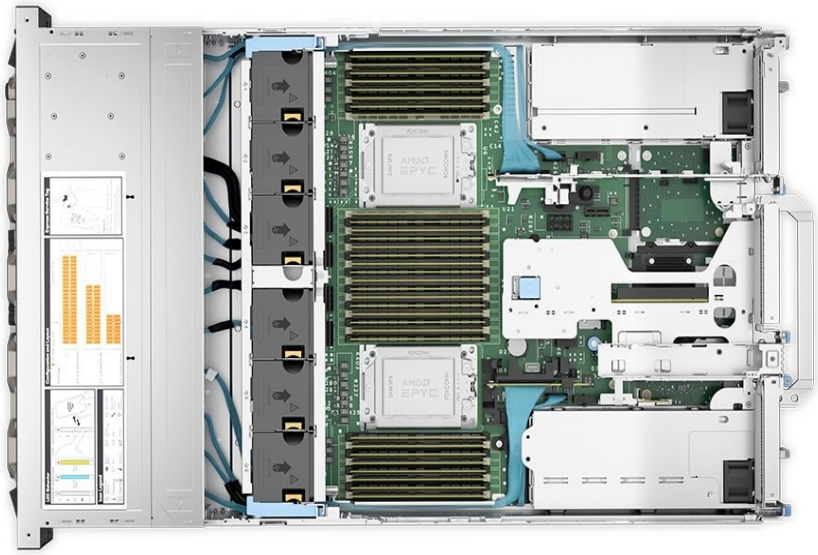
2 VÝBĚR KOMPONENT PRO HPC A GPU CLUSTERY

V první řadě je zapotřebí si říct, že vybudovat HPC cluster si může téměř každý. Stačí k tomu pouze pár počítačů, dostatečně rychlá počítačová síť a správně zvolený a navržený operační systém. Jelikož ale spousta firem nemůže spoléhat na obyčejné počítače, jsou na trhu i výrobci, kteří poskytují profesionální řešení. Každý si tak může sestavit vlastní cluster dle jeho požadavků na výkon.

2.1 Hardware od společnosti Dell

Dell je americká společnost, která se specializuje na výrobu profesionálních výpočetních zařízení, ale i síťových prvků určených primárně do firem a datových center. Mezi další její aktivity patří i výroba běžných počítačů. [2]

Speciálním zařízením určeným právě do výpočetních clusterů patří Dell EMC Power Edge R7525. Server je možné osadit až dvěma procesory řady AMD EPYC 2. generace s podporou až 64 procesorových jader a celkem 512 GB operační paměti. Další výhodou je přítomnost verze PCIe řady 4.0 a podpora řadičů SAS a NVMe pro úložiště dat. Grafickými akcelerátory jsou v tomto případě dvě karty výrobce NVIDIA, model A100, které obsahují Tensor Core jednotky specializované pro operace s umělou inteligencí a strojovým učení. [3][4][5]

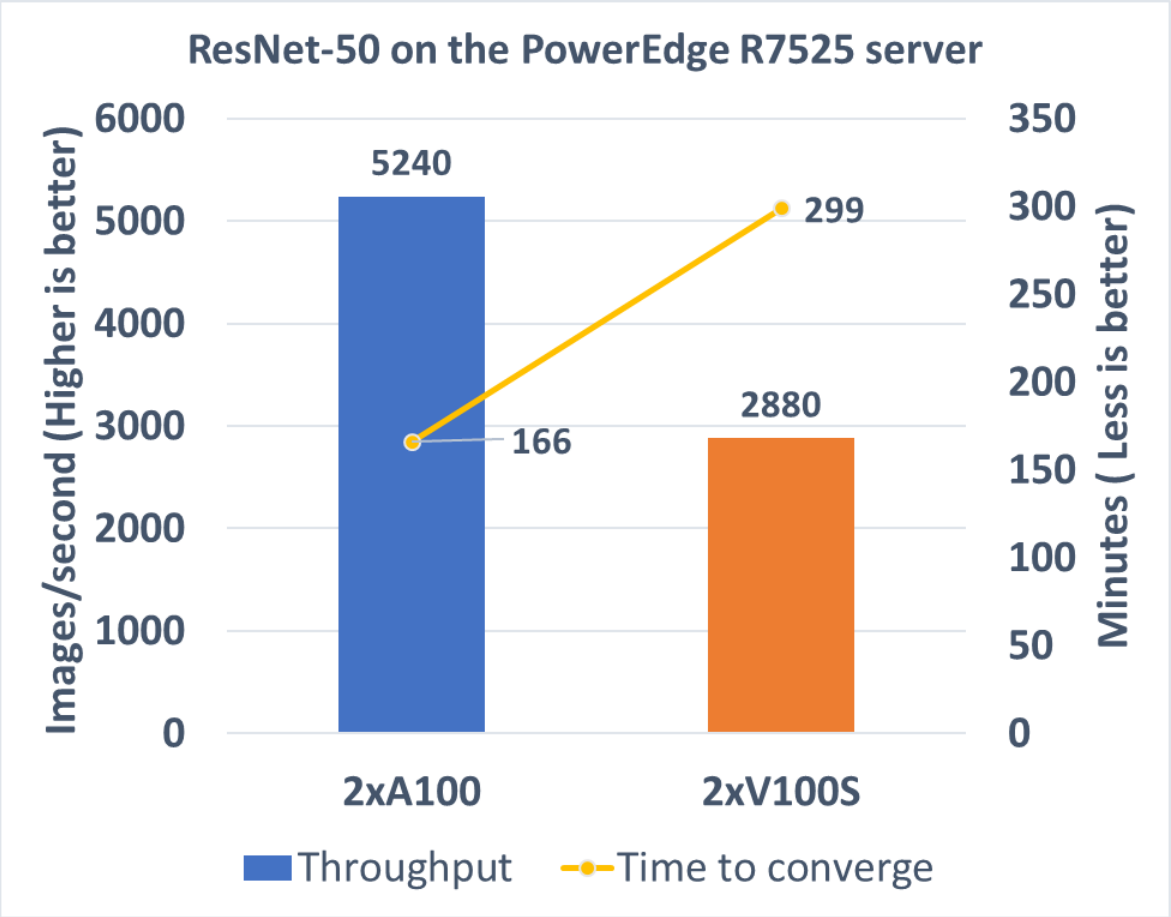


Obrázek 1 Pohled do útrob Power Edge R7525 [5]

2.1.1 Výkonové srovnání grafických karet

Jedním ze způsobů jak porovnávat mezi sebou výkonnost výpočetních stanic je provést jejich tzv. benchmark. Benchmarkem ověříme parametry jednotlivých karet mezi sebou ve zvolených algoritmech. Mezi používaný algoritmus řadíme MLPerf Training, který měří, jak rychle dokáže systém trénovat modely umělé inteligence.

Výkonnostní srovnání bylo provedeno se starší generací karet řady NVIDIA Tesla V100S.

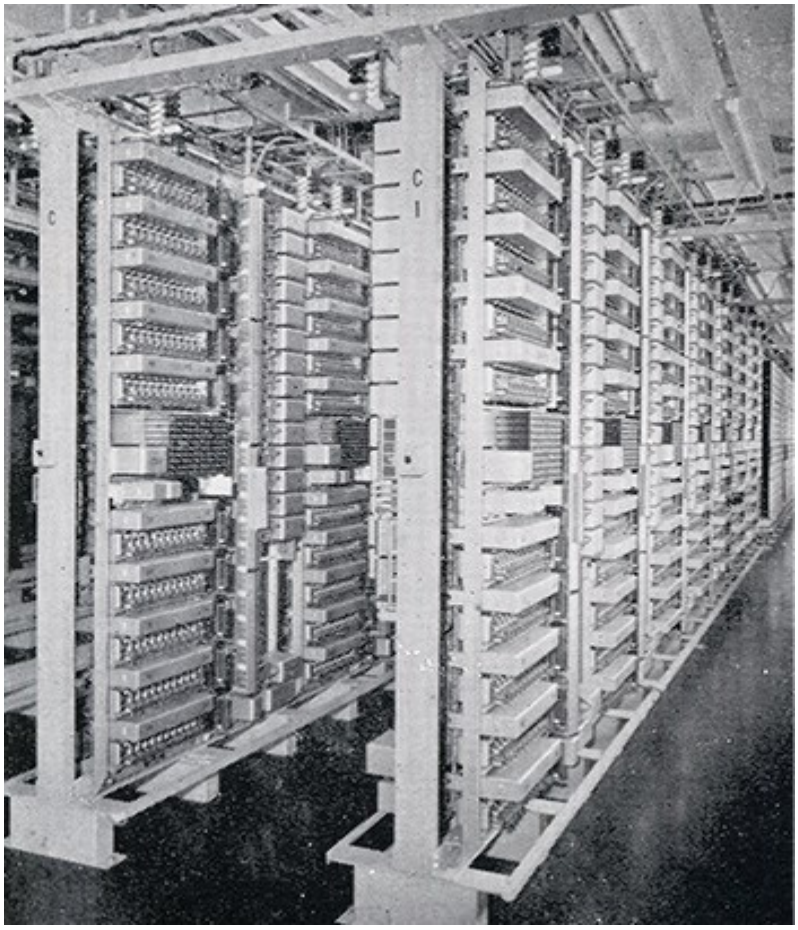


Obrázek 2 Graf srovnání výkonu [4]

Algoritmus měří čas potřebný k trénování sady dat tak, aby byla dosažena alespoň 75,9% přesnost. Z výše uvedeného srovnání vyplývá, že grafické karty řady A100 dokážou zpracovat o více jak 40 % více dat než předchůdce V100S. Stejně tak má nižší požadovaný čas učení o 45 %. [4]

2.2 Hardware od společnosti Fujitsu

Mezi dalšího hojně využívaného výrobce napříč výpočetními servery můžeme zařadit i japonského výrobce Fujitsu. Společnost se stala velmi populární poté, co v roce 1923 zasáhlo silné zemětřesení městečko Kanto, a došlo tak ke zničení telekomunikační infrastruktury měst Tokia a Jokohamy. Tehdy ještě společnost Fuji Electric Co., Ltd. byla oslovena, aby obnovila telefonní ústřednu, ovšem s automatickým přepínacím systémem. [6]



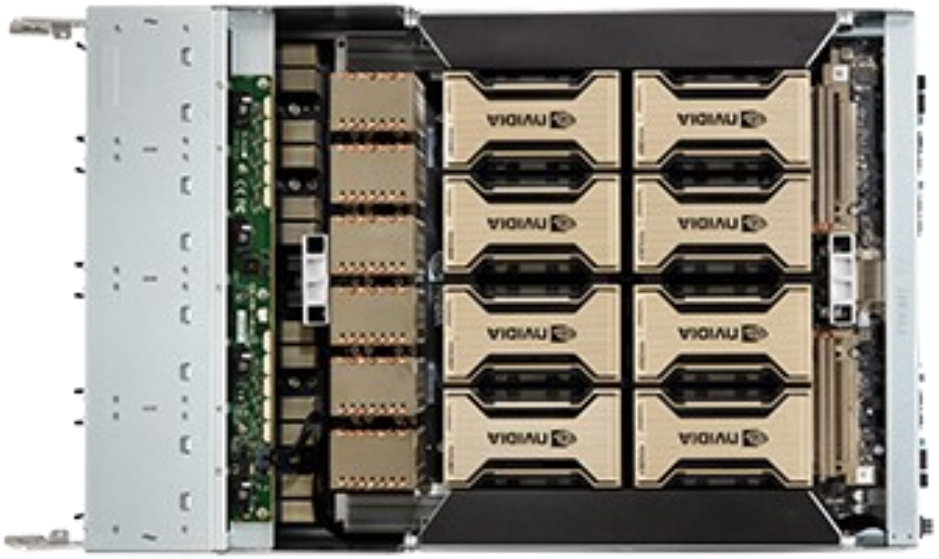
Obrázek 3 Automatický telefonní systém [6]

Z jejich nabídky si můžeme zvolit Fujitsu Server PRIMERGY GX2570 M6. Jde o vysoce výkonný server určený právě pro náročné GPU operace jako je AI. Výrobce navíc na svých stránkách uvádí, že server je certifikovaný přímo společností NVIDIA. [7]

Celý server pak je možné osadit v následující maximální konfiguraci:

Tabulka 1 Hardwarové komponenty serveru PRIMERGY GX2570 M6 [7]

Typ komponenty	*Název*
Procesor	Intel Xeon Gold 53xx, Intel Xeon Gold 63xx
Operační paměť	512 GB – 2 TB
Počet slotů operační paměti	32
Počet PCI-Express 4.0 x16	10
Diskové šachty	6x NVMe/SAS/SATA (vpředu) + 4xNVMe (vzadu)
GPU akcelerátory	až 8x NVIDIA A100 80 GB SXM4
Zdroj	Celkem 4 kusy, max. příkon 3000 W, redundance 2+2
Rozměry	485 x 947 x 175 mm



Obrázek 4 Pohled do útrob PRIMERGY GX2570 M6 [7]

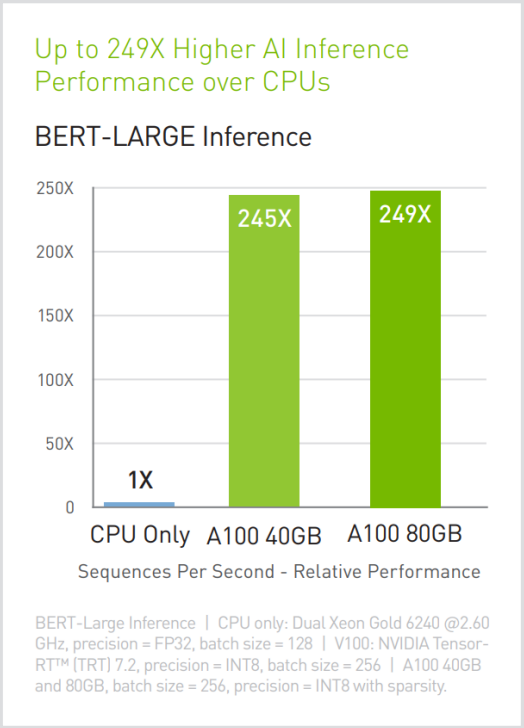
2.2.1 Výkonové srovnání NVIDIA A100 80 GB

NVIDIA A100 je jednou z nejvýkonnějších grafických akcelerátorů aktuálně na trhu. Hodí se právě do pracovních stanic, které pracují s umělou inteligencí, analýzou dat a vysoce náročnými výpočty. Oproti své předchozí generaci NVIDIA Volta poskytuje až 20x vyšší výkon. Mezigenerační skok přinesl dvojnásobně vyšší operační paměť a dvojnásobně vyšší datovou propustnost, která se pohybuje na hranici 2 TB/s.

Mezigeneračně dochází k nárůstu výkonu oproti staršímu modelu NVIDIA Tesla V100. Model A100 80 GB vykazuje 3x vyšší výkon v trénování AI oproti svému předchůdci. [8]

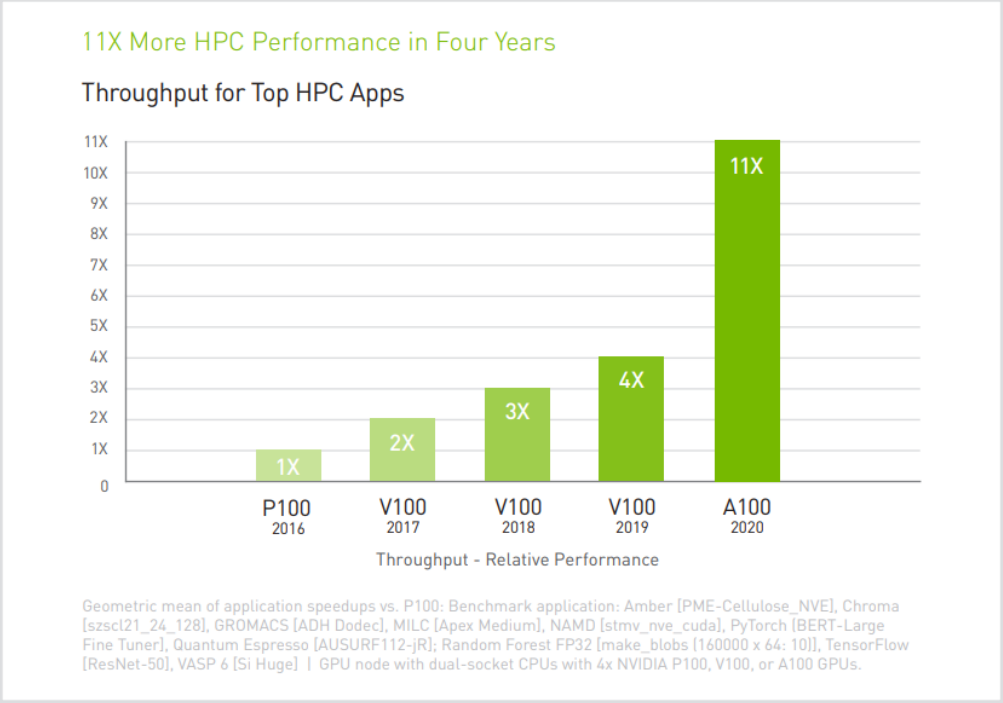
Obrázek 5 Srovnání grafických akcelerátorů [8]

Stejně tak je až 249x výkonnější než procesor řady Dual Xeon Gold 6240. To nám ukazuje jednoznačně, proč se právě pro strojové učení využívají hlavně grafické akcelerátory.



Obrázek 6 Srovnání výkonu s procesorem [8]

Poslední z grafů nám jasně říká, že během čtyřletého vývoje dosáhla 11x tak vyššího výkonu v HPC oproti starší generaci NVIDIA Tesla P100 z roku 2016 a 4x vyššího výkonu oproti svému předchůdci NVIDIA Tesla V100 z roku 2019.



Obrázek 7 Mezigenerační srovnání řadami NVIDIA [8]

2.3 Návrh stanice z běžně dostupného hardwaru

Pro domácí použití, případně do menších firem, se nemusí vyplatit koupě profesionální stanice od výrobců, jelikož cena je někdy až astronomická a výpočetní výkon je obrovský. Navíc řada menších firem nedisponuje takovým rozpočtem, aby si mohla takový nákup dovolit. Máme však možnost si takovou stanici sami vytvořit na míru našim potřebám a s požadovaným výkonem. Jestliže nám výkon již nebude dostačovat, jednoduše vyměníme slabší hardware za výkonnější.

2.3.1 Počítačová skříň

Do návrhu stanice jsem zvolil takovou počítačovou skříň, do které je možné osadit až 8 grafických karet, základní desku a zdroj. Skříň je o standartním rozměru 4U, je tedy vhodné ji nainstalovat i do rackových skříní. Konstrukce je vyrobena z kvalitního plechu a působí velmi profesionálně. Uvnitř nechybí šestice výkonných ventilátorů, které zajistí vysoký průtok vzduchu, a nebude tak hrozit přehřátí hardware. Na čelní straně najdeme tlačítka pro vypnutí či restartování spolu s dvojicí USB 2.0 konektorů a LED diody indikující zapnutí stanice či vytížení pevného disku. [9]

Obrázek 8 Počítačová skříň [9]

2.3.2 Základní deska

V návrhu jsem zvolil desku od výrobce ASRock, model Taichy X399 s chipsetem AMD X399. Deska disponuje celkem 4 PCIe sloty verze 3.0, z nichž každý ze slotů má plných 16 linek. Díky takovému množství linek je možné využít plný výkon grafických karet, ovšem v případě zapojení všech čtyřech možných karet deska sama rozdělí první a třetí slot s podporou 16 linek a druhý se čtvrtým na 8 linek. Nechybí ani podpora ECC[1] pamětí, díky kterým je možné ji používat v serverových řešeních. Další parametry si můžeme ověřit v následující tabulce: [10]

Tabulka 2 Další parametry základní desky [10]

Funkcionalita	*Text*
Podpora procesorů	AMD TR4 socket pro procesory řady Ryzen či Threadripper
Operační paměť	Čtyř kanálové DDR4 s podporou ECC, maximální kapacita 128 GB
Konektivita	Dvojice LAN Gigabit portů, podpora bezdrátové sítě WiFi
Sloty	4x PCI Express 3.0 x16 (při zapojení dvou grafických karet)
	1x PCI Express 2.0 x1
Úložiště	8x SATA3 6.0 Gb/s s podporou RAID[2] diskového pole
	3x M.2 socket pro SSD disky

[1] ECC – technologie umožňující detekci a opravu chyb při přenosu dat
[2] RAID – Redundant Array of Inexpensive Disks – technologie umožňující ukládat data na více disků a navzájem tak zálohovat jejich obsah

Obrázek 9 Použitá základní deska [10]

2.3.3 Procesor

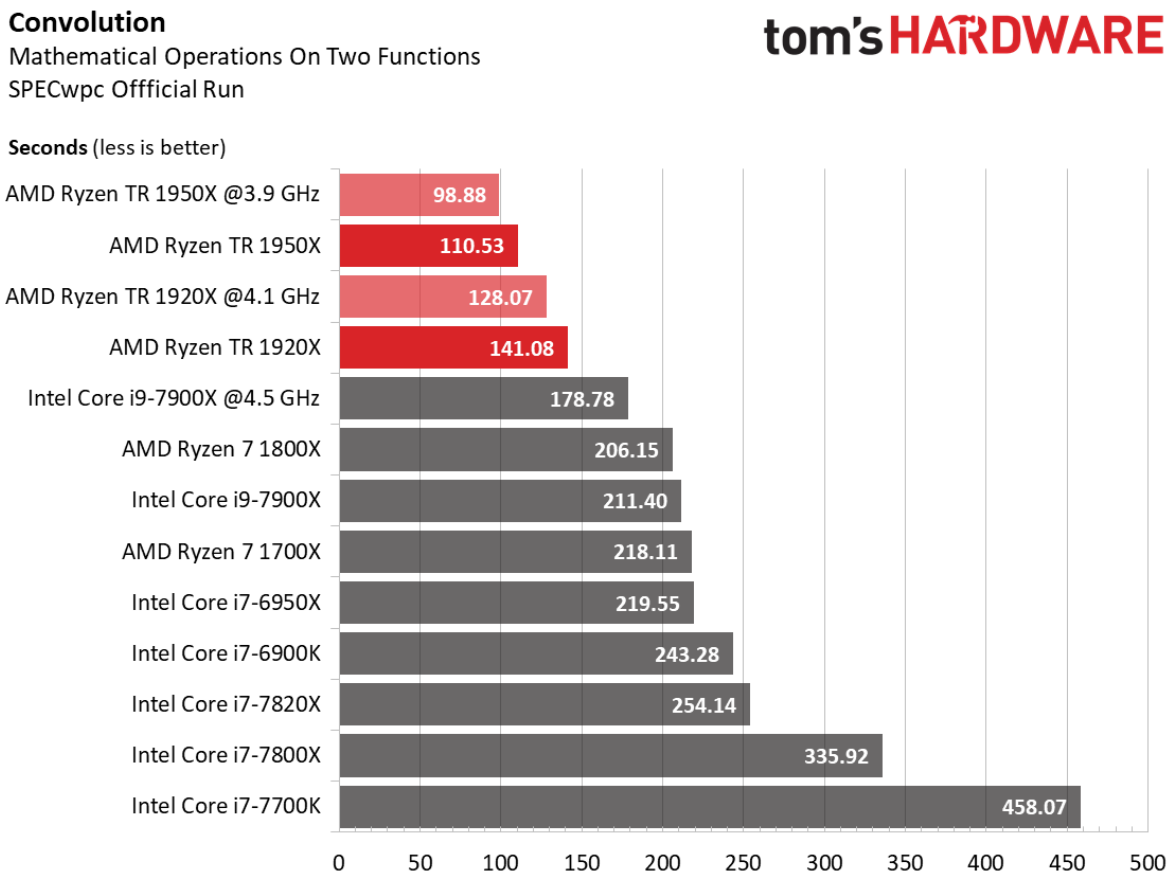
Optimálním poměrem mezi cenou a výkonem může být procesor od AMD, a to model Threadripper 1920X. Procesor má k dispozici celkem 12 fyzických jader, jež je ale schopen rozdělit na 24 vláken. Díky tomuto dělení je umožněno spouštět až dva procesy na jednom jádru a zrychlit tak běh aplikací. Frekvence procesoru je v základu nastavena na vysokých 3.5 Ghz s možností automatického zvýšení až na 4.0 Ghz. Další výhodou tohoto procesoru je schopnost vlastního zvyšování jeho frekvence, což nám umožní zvýšení jeho výkonu. Procesor neobsahuje grafické jádro, tudíž je zapotřebí jej připojovat vždy s nějakou grafickou kartou kvůli zajištění obrazového výstupu ze stanice. V neposlední řadě je podpora virtualizace přímo v procesoru. [11]

Další informace zjistíme z níže přiložené tabulky:

Tabulka 3 Další parametry procesoru [12]

Parametr	*Hodnota*
Produktová řada	AMD Ryzen Threadripper Procesor
Výrobní technologie	14nm tranzistory
Kapacita L1 cache	1,125MB
Kapacita L2, L3 cache	6 MB, 32 MB
Podporované operační systémy	Windows 10, RHEL, Ubuntu

Procesor se hodí do poloprofesionálních pracovních stanic, kde je důležitá nízká pořizovací cena, ovšem ne na úkor markantního snížení výkonu. Nabízí tedy přesně takový výkon, který je konkurenceschopný s konkurenčními procesory od firmy Intel za srovnatelnou cenu.



Obrázek 10 Výkonové srovnání s konkurencí [13]

Srovnání můžeme vidět na přiloženém grafu výše, kde byla použita sada benchmarků SPECwpc pro pracovní stanice s širokým množství využití. Benchmark provádí různé řady matematických výpočtů optimalizovaných pro paralelizaci. Skládá se z operací založených na dvou funkcích, jejichž výsledkem je třetí funkce. Takové benchmarky využívají hlavně datovou propustnost a šířku datové sběrnice mezi pamětí a procesorem. Přední příčky na grafu zabírají hned čtyři procesory od AMD a nechávají tak konkurenci od Intelu za sebou s přehledným náskokem. [13]

2.3.4 Operační paměť

Operační paměť hraje v návrhu výkonných pracovních stanic významnou roli, jelikož u prováděných výpočtů velice záleží na přesnosti práce s daty. Větší systémy se většinou řídí pravidlem, kdy na jedno výpočetní jádro procesoru připadají 2–3 GB operační paměti. Při návrhu pracovní stanice ovšem musíme pracovat s ohledem na využití takového zařízení a s požadavky budoucích uživatelů a vývojářů aplikací. Některé aplikace v oblasti náročných matematických výpočtů mohou být velice náročné na kapacitu operační paměti. [14]

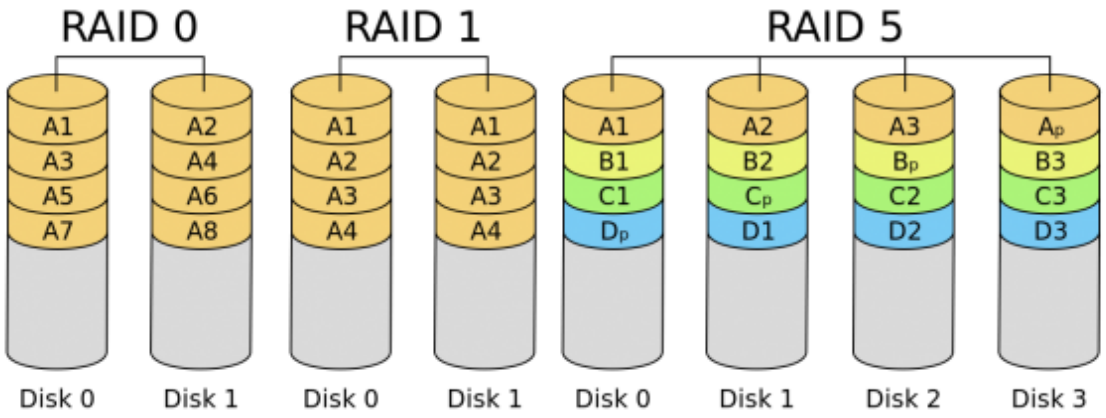
V případě návrhu systémů, na které je kladen důraz na spolehlivost práce s daty, je zapotřebí mít paměti s technologii zvanou Error Correction Code, neboli Kód opravující chyby. Dokáže detekovat a opravit jednobitové či vícebitové chyby, ale také jednobitové chyby opravovat za chodu aplikace. K zajištění ochrany používá celkem 7 bitů k ochraně 32bitových bloků dat, případně 8 bitů k ochraně 64bitových bloků. Bohužel, ne každý paměťový modul zvládne s touto technologií pracovat a ani ne každá základní deska. Dnes je již téměř standardem ve všech profesionálních serverech. [15]

Vzhledem k použitému procesoru jsem se rozhodl později ve své pracovní stanici použít operační paměti s celkovou kapacitou 32 GB.

2.3.5 Diskové úložiště

Při návrhu, jak pracovat s daty na naší stanici, bychom se měli primárně snažit o to, aby nedošlo ke ztrátě dat v případě poruše disku. Takovou situaci jsme schopni řešit technologií zvanou Redundant Array of Inexpensive Disks, která nám umožní za použití více disků vytvořit jeden úložný disk, aniž bychom se museli obávat ztráty dat. Principem je ukládání dat na více různých disků, kde v případě selhání jakéhokoliv z nich jsme schopni uložená data dostat zpět neporušená. Mezi základní typy diskových polí řadíme:

- RAID 0 (striping) – základní pole, které ovšem neposkytuje ochranu proti ztrátě dat, ale pouze zvyšuje výkon disků díky zvýšení rychlosti čtení a zápisu.
- RAID 1 (mirroring) – je zapotřebí alespoň dvou disků, kde každý z disků ukládá stejná data. Nevýhodou může být, že vždy musíme obětovat navíc jeden disk s totožnou kapacitou.
- RAID 5 (single parity) – u tohoto typu diskového pole je zapotřebí alespoň tří disků, kde úložný prostor jednoho disku zabírají samoopravné kódy. Nevýhodou může být pomalejší rychlost zápisu dat, jelikož musí dojít k vypočítání a uložení samosprávných kódů.



Obrázek 11 Ukládání dat na jednotlivá disková pole, kde A, B, C jsou data a D_p, C_p, B_p či A_p paritní data [16]

Naše stanice bude pracovat s celkem dvěma disky připojenými přímo do M.2 slotu na základní desce. Ochrana dat bude zajištěna díky RAID 1. Tento typ RAIDu nám bude vyhovovat nejvíce, jelikož nám plně dostačuje pouze ochrana proti ztrátě dat v případě, že bude jeden z disků z nějakého důvodu poškozen. [16][17]

2.3.6 Grafické karty

Mezi světové výrobce se řadí společnost NVIDIA a její grafické jednotky speciálně určené pro provoz v datových centrech a výpočetních serverech. Tato společnost rovněž stojí za technologií zvanou CUDA. CUDA je obdobou procesorového jádra ovšem s tím rozdílem, že je méně sofistikované, ale za to v mnohem větším počtu. Běžné procesory dosahují dvou až třiceti dvou jader, ale jádra CUDA se počítají na stovky, u výkonnějších karet na tisíce. Je to do jisté míry dáno i tím, že procesor musí zvládat zpracovávat více univerzálních úloh oproti grafickým kartám. [18]

Z běžně dostupných grafických karet určených hlavně pro hráče počítačových her máme aktuálně (duben 2022) k dispozici modelovou řadu RTX 30xx, která obsahuje celkem 9 modelů z nichž se celkem 4 řadí do vyšší třídy. Jsou to převážně tyto modely:

Tabulka 4 Srovnání čtyř nejvyšších modelů řady RTX 30xx [19]

Model	*RTX 3090*	*RTX 3080Ti*	*RTX 3070Ti*	*RTX 3060Ti*
Počet CUDA	10496	10240	6144	4864
Frekvence jádra (Ghz)	1,70	1,67	1,77	1,67
Velikost paměti (GB)	24	12	8	8

Můžeme si všimnout, že nejvyšším počtem zvýšení CUDA jader dochází mezi RTX 3070Ti a RTX 3080Ti, kde rovněž dochází i ke zvýšení velikosti paměti o 4 GB.

Obrázek 12 Ukázka grafické karty RTX 3090 [20]

Pro profesionály má NVIDIA připravenou řadu RTX A Series, kde má k dispozici více než dvanáct karet. Dělí je zároveň na dva segmenty – karty určené do pracovních stanic a do serverů. Oproti kartám pro hráče počítačových her se liší například kapacitou operační paměti či zabudovanou podporou ECC na pamětech.

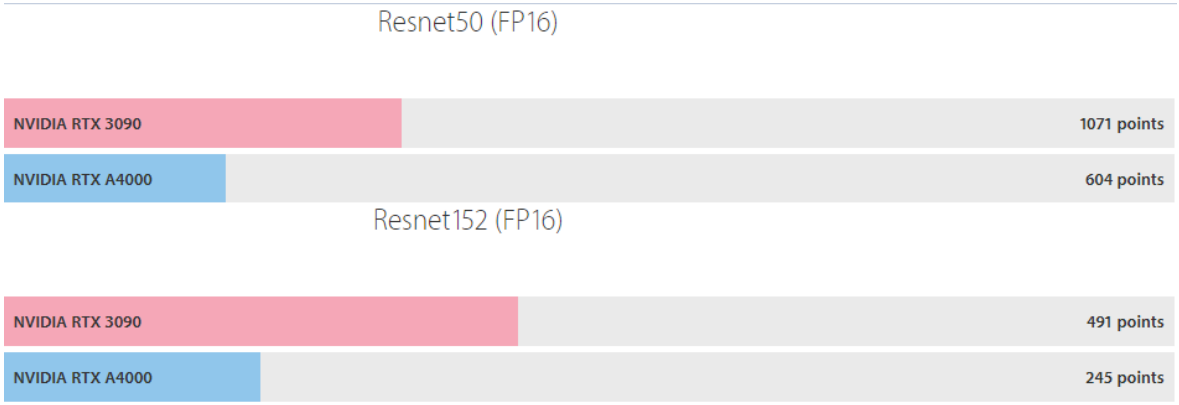
Zároveň jsou tato zařízení stavěné pro provoz v serverech a jsou tak i navrhovány. Oproti herním kartám zabírají pouze jeden PCI slot, můžeme jich tudíž zapojit více vedle sebe.

Mezi karty určené do pracovních stanic řadíme například následující modely:

Tabulka 5 Srovnání grafických karet pro pracovní stanice [21]

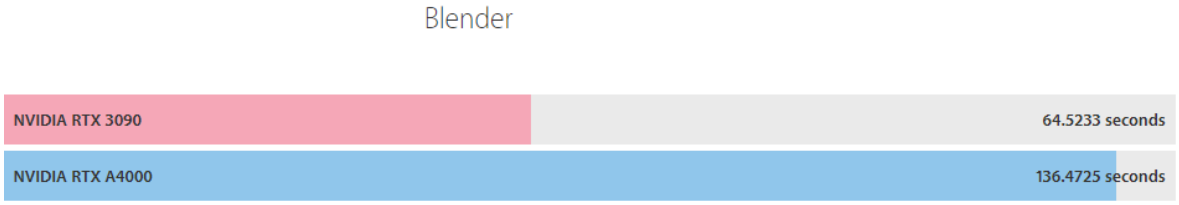
Model	*RTX A2000*	*RTX A4000*	*RTX A5000*	*RTX A6000*
Počet CUDA	3 328	6 144	8 192	10 752
Velikost paměti (GB)	12	16	24	48

V návrhu naší pracovní stanice použijeme jednu z nejvyšších karet v nabídce z řady RTX 3090 pro běžné hráče, protože dle výkonových srovnání dosahuje opravdu vysokého výkonu a je v době psaní této práce k dispozici. Z profesionální sféry použijeme naopak jednu ze střední kategorie, konkrétně RTX A4000, rovněž z důvodu aktuální dostupnosti na trhu. Níže se můžeme podívat na srovnání těchto dvou karet, co se týče jejich výkonu v různých benchmarcích:



Obrázek 13 Srovnání výkonu pomocí benchmarku Resnet [22]

ResNet je zkratka pro Residual Network, což je umělá neuronová síť. Označení ResNet50 znamená základní ResNet, který ovšem může pracovat s 50 vrstvami neuronové sítě. V případě ResNet152 je to až celkem 152 neuronových sítí. Z benchmarku je patrné, že RTX 3090 dosahuje téměř jednou tak vyššího výkonu než v případě RTX A4000. [23]



Obrázek 14 Srovnání výkonu v grafickém benchmarku [22]

Dalším srovnáním může být benchmark ve velice známém softwaru na tvorbu grafických animací a různých scén. Benchmark měří dobu, za kterou grafická karta zvládne vyrendero-vat obrázek. I zde má RTX 3090 značně navrch oproti RTX A4000.

2.3.6.1 Nevýhoda RTX 3090

Mezi nevýhody karty řady RTX 3090 můžeme zařadit jejich poměrně vysokou spotřebu elektrické energie. Tu nám vyjadřuje tzv. Thermal Desing Power, zkráceně TDP. Jedná se o maximální tepelný výkon, který musí být chladič schopný odvést z čipu. U zařízení jako je procesor či právě grafická karta, udává tato hodnota maximální spotřebu elektrické energie bez zásahu do nastavení. Karta RTX 3090 má hodnotu TDP 350 wattů, což má za následek při špatně navrženém chlazení špatný odvod tepla a přehřívání. Naproti tomu karta RTX A4000 má hodnotu TDP o dost menší, rovných 140 wattů. [24][25]

3 MOŽNOSTI MONITORINGU CLUSTERŮ

3.1 Motivace pro monitoring

Veškerá výpočetní zařízení, ať už profesionální servery či vlastní navržené zařízení se skládají z počítačového hardwaru, který ovšem není bezchybný. Může se jednat o chyby hardwaru (například poškození grafické karty, poškození chlazení), nebo o chyby softwarové (zaseknutí operačního systémů, pád programu…). Proto je potřeba veškeré tyto komponenty a systémy monitorovat tak, aby administrátor celého systému měl přehled o tom, co se s jeho zařízením děje, případně jestli zákazníkovi, který si objednal služby výpočetního clusteru funguje vše tak, jak má. Profesionální výrobci používají vlastní monitorovací systémy, které poskytují administrátorovi veškerá data o stavu serveru. Výrobce Fujitsu má systém zvaný Remote Desktop Controller. Co se týče výrobce Dell, ten má rovněž vlastní řešení, které nazývá Integrated Dell Remote Access Controller, zkráceně iDRAC. [26][27]

3.2 Monitorované parametry

Každá komponenta našeho serveru by měla být monitorována. Administrátor systému tak bude mít přehled o stavu hardwaru či softwaru. Mezi parametry, které chceme monitorovat, zahrnujeme například:

- Dostupnost – základní parametr, který nám říká, zda je server dostupný či nedostupný.
- Otáčky ventilátorů – schopnost monitorovat a nastavovat otáčky ventilátorů, které zajišťují optimální průtok vzduchu přes komponenty.
- Využití procesorových jader – při běhu více aplikací na jedné stanici monitorujeme vytížení každého výpočetního jádra. Vytížení se udává v procentech.
- Monitoring procesů – umožňuje nám zobrazit stavy jednotlivých procesů v systémů a zjišťovat, které procesy jsou narušeny.
- Využití operační paměti – zobrazení aktuální maximální kapacity operační paměti a její zaplnění.
- Stav disků – zobrazení disků, jejich zaplnění, životnost, teplota, stav diskového pole.
- Stav grafických karet – detekce, zda je karta zapojena či odpojena, sledování zatížení, teploty, spotřeby elektrické energie.
- Vytížení sítě – sledování stavu síťového rozhraní serveru, datový tok paketů, vytížení přenosové kapacity.

- Přístupy – logování přihlašování do systému, aktivit a připojení externích zařízení. [28]

3.3 Monitorovací systémy

Na trhu existuje spousta řešení, které nám dokáží monitorovat všechny aktivity našeho serveru. Některá řešení jsou zpoplatněna, ovšem některá ne. Úplně tedy nedává smysl vyvíjet vlastní aplikaci na monitorování.

3.3.1 Zabbix

Firma Zabbix LLC vznikla v roce 2005 v Lotyšsku. Věnuje se vývoji monitorovacího softwaru, který monitoruje veškeré dění na síti, fyzických serverech či virtuálních strojích. Software je plně bezplatný. [29]

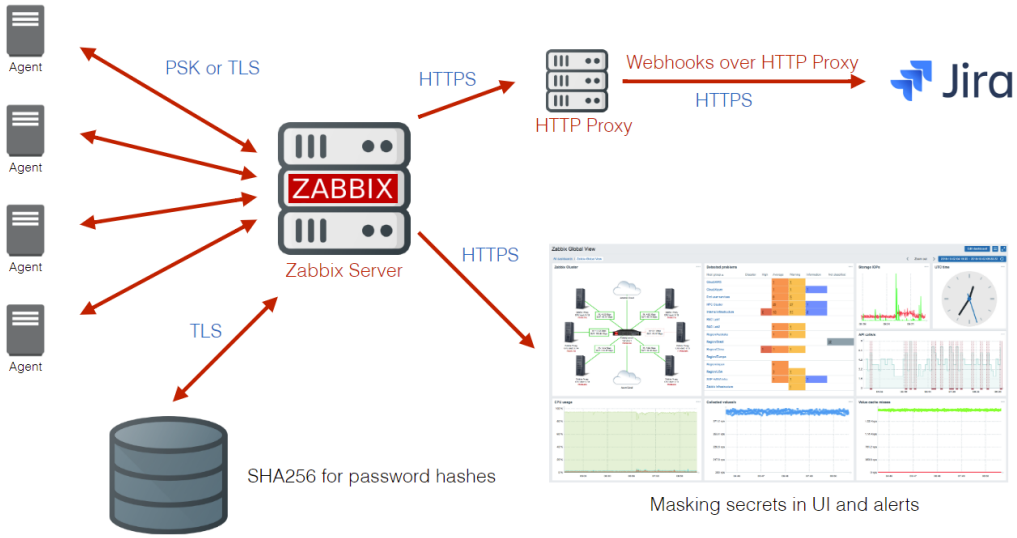
Mezi jeho hlavní funkce zahrnujeme:

- Sbírání dat – kontrola dostupnosti a výkonu, podpora SNMP[3], shromažďování dat v různých intervalech.
- Definice vlastního chování – umožňuje nastavit mezní parametry z monitorovacích dat, kdy monitorované hodnoty představují potencionální problém.
- Široká nabídka možností upozornění – zaslání upozornění pomocí emailu, SMS či jiných komunikačních nástrojů jako je například Slack, MS Teams, Telegram atd…
- Grafy v reálném čase – veškeré monitorované parametry jsou vykreslovány v reálném čase.
- Ukládání dat – logování dat do databáze pro pozdější analýzy.
- Webové rozhraní – na veškerá data je možné nahlížet z grafického webového rozhraní.
- Zabbix API – integrace jiných softwaru pro analýzu, případně ovládání.

Zabbix se skládá z několika hlavních komponent, které jsou nepostradatelné pro správný chod celého systému. Řadíme mezi ně:

[3] SNMP – Simple Network Management Protocol – sada síťových protokolů pro správu počítačové sítě a její analýzu

- Server – hlavní komponenta, která sbírá veškerá data a informace ze serverů, ukládá nastavení.
- Databáze – uložení nasbíraných dat.
- Webové rozhraní – administrátorovi poskytuje data v reálném čase spolu s grafy a různými přehledy.
- Proxy server – volitelná součást, která může sloužit k rozložení zátěže v případě monitoringu rozsáhlé sítě. [30]



Obrázek 15 Funkční schéma systému [31]

3.3.2 Dell iDRAC

Téměř všichni výrobci profesionálních pracovních stanic mají vlastní monitorovací software, který se stará o monitoring veškerých prvků v serveru. Jde o velice sofistikované a specifické řešení, které je ovšem velice spolehlivé. Má nespočetně výhod, mezi něž patří například monitoring serveru i při vypnutém stavu, včasné varování administrátora před hrozícími skutečnostmi, jakými jsou např. porucha disk či chyba v diskovém poli.

iDRAC je software společnosti Dell, který má za úkol monitorovat, vyhodnocovat a ukládat data o serverech a informovat administrátora o potencionálních hrozbách. Systém je zabudovaný v základní desce serveru. Instalovaný systém tedy nemá vliv na funkci monitoringu. Další velice užitečnou funkcionalitou je i možnost server vzdáleně ovládat pomocí konzole bez nutnosti být fyzicky u serveru přítomen. Administrátor tak může instalovat operační systémy či provádět aktualizace firmware vzdáleně. [32]

iDRAC je dostupný v následujících variantách pro různé velikosti datových center:

- iDRAC Basic
- iDRAC Express
- iDRAC Enterprise
- iDRAC Datacenter

Mezi výhody takového systému můžeme zařadit také zabezpečení přístupu ke vzdáleným serverům, kdy správci mohou na dálku provádět různá nastavení, aniž by došlo k ohrožení bezpečnosti serveru i sítě.

3.3.2.1 Monitorovaná data

Systém monitoruje veškeré komponenty, které jsou zapojené na základní desce. Stejně tak je schopen monitorovat i teploty jednotlivých čipů integrovaných na základní desce, jako je například chipset. Ze základních monitorovací dat stojí za zmínění následující monitoringy:

- Zobrazení stavu serveru, zda je zapnutý, vypnutý či zaseknutý.
- Stav síťových adaptérů, jejich technická data, maximální datová propustnost.
- Teplotní senzory umístěné na základní desce pro monitoring teploty čipů.
- Vytížení procesoru, jeho teploty, napětí, počet výpočetních jader, technická data.
- Monitoring operační paměti, její maximální kapacitu, pracovní napětí, technická data.
- Monitorování spotřeby elektrické energie.

3.3.2.2 Správa serveru na dálku

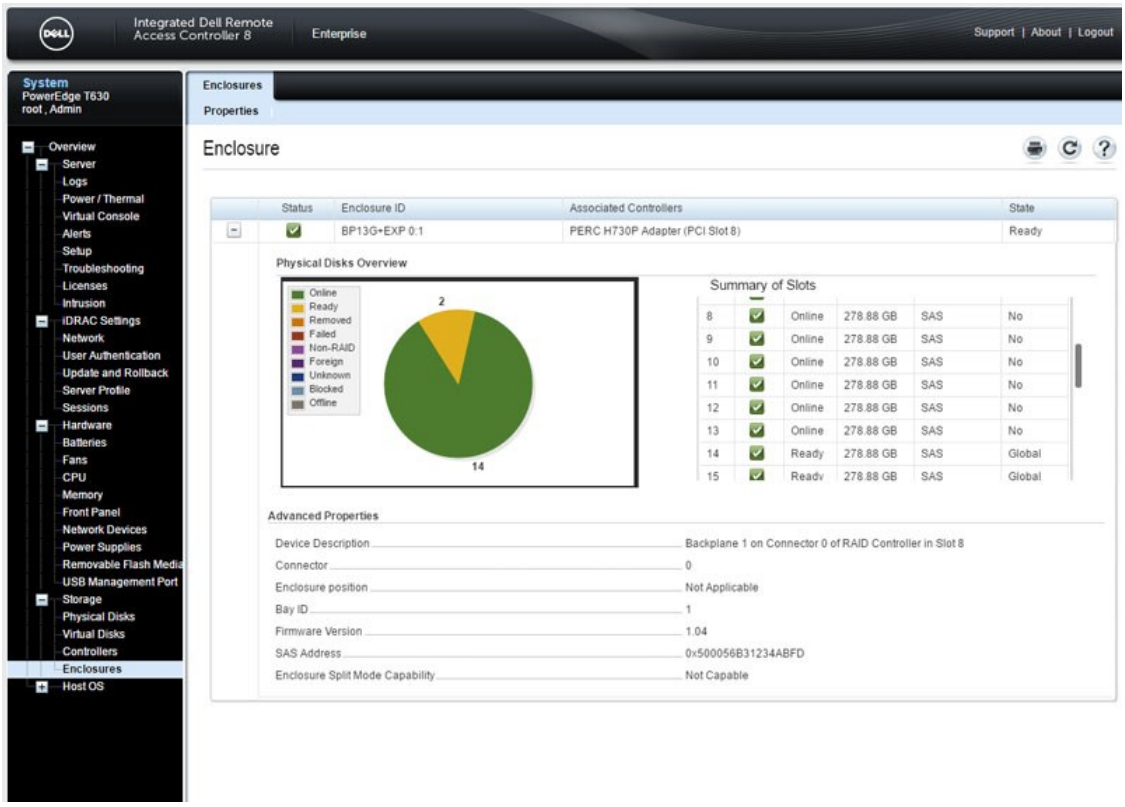
iDRAC neslouží pouze a jen k monitorování, ale také umožňuje provádět různá nastavení podle potřeby administrátora. Za zmínku stojí určitě:

- Konfigurace informačního panelu na serveru (pokud je vybaven).
- Virtuální konzole pro správu hypervizoru či virtuálních systémů.
- Vzdálená instalace operačních systémů.
- Nastavení a udržování diskových polí, jejich vytváření, úpravy, detekce nových disků, šifrování.
- Konfigurace připojených řadičů.

3.3.2.3 Zabezpečení připojení

Důraz musí být kladen i na bezpečnost přenášených dat mezi pracovní stanicí administrátora a samotným serverem. Proto systém iDRAC obsahuje velkou řadu možností, jakým způsobem může být zabezpečen samotný přenos informací či aktualizací.

- Tvorba a správa vlastních SSL certifikátů.
- Podepisování aktualizací firmwaru.
- Ověřování uživatelů za pomocí LDAP.
- Dvoufaktorové ověřování pomocí chytrých karet.
- Nastavení oprávnění napříč administrátory.
- Zabezpečený přístup pomocí zabezpečeného přístupového terminálu.
- Blokování přihlašování do systému v případě špatně zadaných údajů.
- Omezení přístupů ze specifických IP adres. [33]

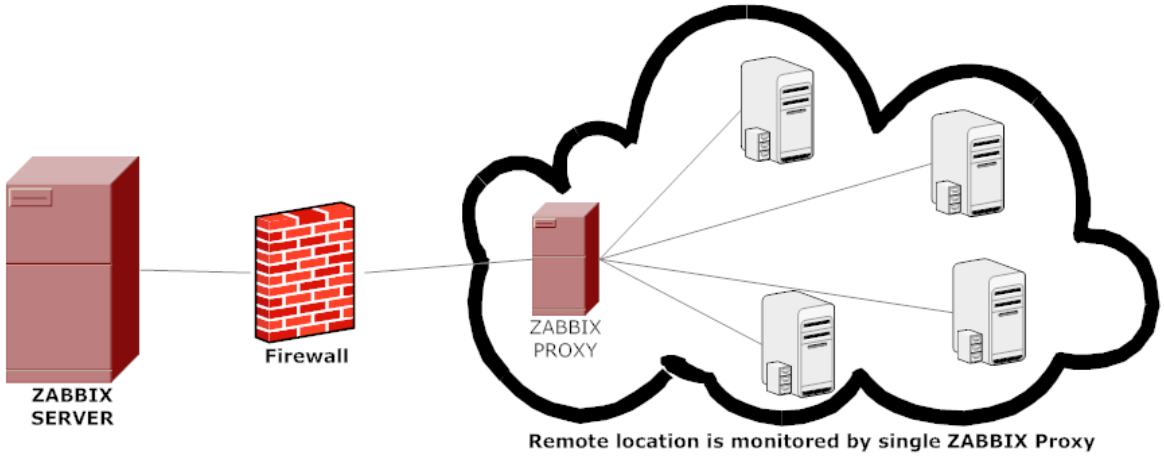


Obrázek 16 Náhled na prostředí systému iDRAC [34]

3.4 Centrální monitoring více serverů na jednom rozhraní

Výše popsané monitoringy jsme si vysvětlili na jednotlivých serverech. Dozorovat každý server jednotlivě je velice nepraktické, jelikož bychom museli vždy hlídat data každého serveru samostatně, což by bylo při větším množství zařízení poměrně náročné. Proto se nám nabízí řešení centralizace do jedné webové stránky či systému, odkud budeme monitorovat všechny servery z jednoho místa.

Zabbix nám nabízí využít svou proxy komponentu, díky které dokážeme připojit více jednotlivých stanic, které se budou hlásit do hlavního serveru, odkud můžeme data vyhodnocovat, zobrazovat a zpracovávat. Takovýmto způsobem dokážeme monitorovat stovky serverů.



Obrázek 17 Princip centralizovaného monitoringu Zabbix [35]

Zabbix je natolik komplexní, že dokáže pracovat i s daty, které shromažďuje rozhraní iDRAC. Můžeme tedy kombinovat jak data z vlastních výpočetních serverů, tak i těch profesionálních. Komunikace probíhá pomocí SNMP protokolu. [36]

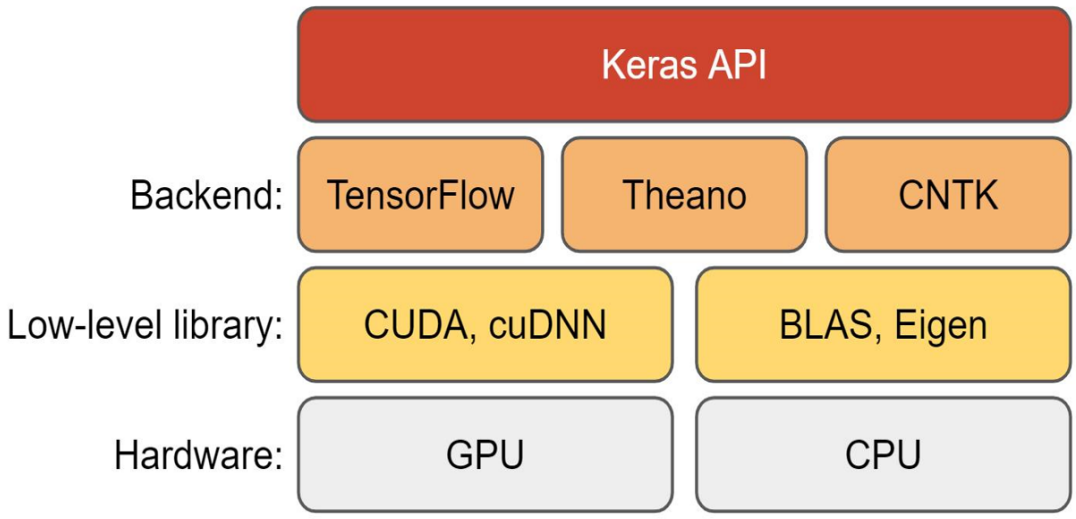
4 NÁVRH SYSTÉMŮ

4.1 Keras

Keras je framework napsaný v jazyce Python, který je užíván pro hluboké učení. Původním záměrem vývoje bylo vytvořit rychlý nástroj, díky kterému bude moci uživatel rychle a jednoduše experimentovat. Využívá především nástrojů TensorFlow. Mezi jeho hlavní přednosti řadíme:

- Jednoduchost pro vývojáře.
- Podpora CPU i GPU.
- Podpora konvoluční a rekurentní sítě a jejich kombinace.

Keras je distribuován pod licencí MIT, tudíž může být použit volně kýmkoliv, dokonce i v komerčních aplikacích.



Obrázek 18 Schéma jednotlivých vrstev [37]

Framework může být instalovaný na systémy podporující TensorFlow verze 2. Zároveň je ale nutné počítat s podporou Pythonu verze 3.6 a výše. Je tedy kompatibilní se systémy:

- Ubuntu verze 16.04 nebo novější.
- Windows 7 nebo novější.
- macOS 10.12.6 (Sierra) nebo novější. [38]

4.2 TensorFlow

Je open source framework určený pro hluboké učení, které vyvinul tým Google Brain v roce 2011. Využívá grafů datových toků, díky kterým reprezentuje jednotlivé výpočty. Jeho přednosti jsou především:

- Velice rychlý na grafických kartách NVIDIA.
- Rozsáhlá podpora clusterizace.
- Flexibilní řešení – obsahuje výkonné matematické operace.
- Obsahuje nástroj TensorBoard, díky kterému můžeme graficky analyzovat vývoj našich modelů.
- Rozsáhlá komunita.
- Široká adopce velkých firem jako je Google, Intel či eBay.

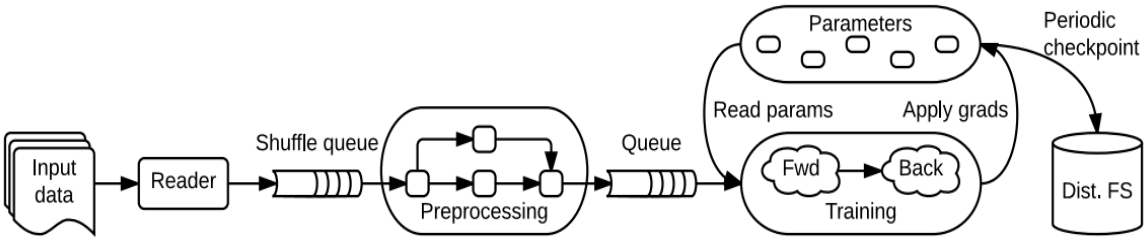
Jak již bylo jedno zmíněno, TensorFlow pracuje s CUDA jádry, které do svých grafických karet implementuje výrobce NVIDIA. Každá grafická karta má různé počty těchto jader a různé kapacity paměti. [39]

TensorFlow může být nasazen na 64bitových systémech:

- Ubuntu 16.04 nebo novější.
- Windows 7 nebo novější (s C++ knihovnami).
- macOS 10.12.6 (Sierra) nebo novější, ovšem bez podpory grafické karty.

Rovněž je zahrnuta podpora Pythonu verze 3.7 a výše. [40]

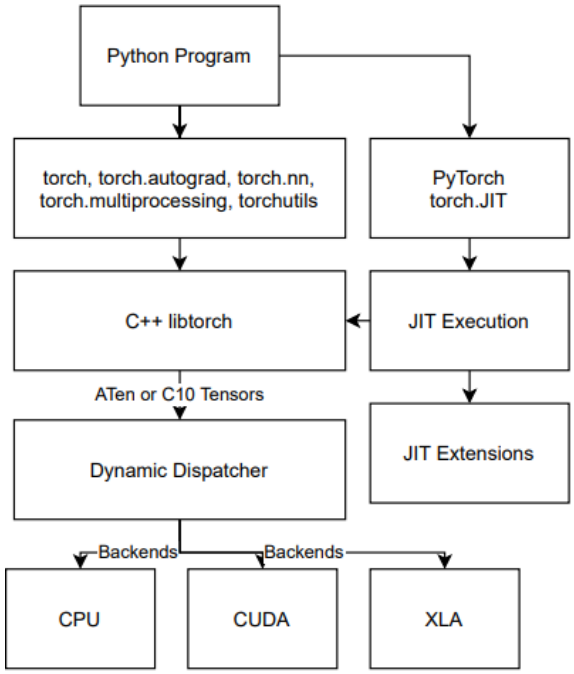
Co se týče podpory grafických karet, zde je potřeba, aby byly použity karty výrobce NVIDIA s architekturou CUDA 3.5, 5.0 6.0, 7.0, 7.5, 8.0 a vyšší. Řadíme sem karty například řady A100, A40, RTX A4000 – A6000 či herní řadu GeForce RTX 3090, 3080, 3070 a 3060. [41]



Obrázek 19 Graf toku dat TensorFlow pro tréninkovou pipeline [42]

4.3 PyTorch

PyTorch je rovněž framework určený k učení neuronových sítí využívající jazyk Python. Jedná se o velice oblíbený nástroj kvůli jeho jednoduchosti a použití. Využívá dynamické výpočty grafů, díky kterým může dosahovat větší flexibility při tvorbě náročnějších sítí oproti svým konkurenčním řešením. Ve srovnání s TensorFlow je zaměřený více na výzkum, je použitelnější díky využití jazyka Python a lépe se učí. Rovněž jako své konkurenční řešení umožňuje provádět operace jak na procesorech, tak na grafických kartách. [43]



Obrázek 20 Ukázka architektury PyTorch [44]

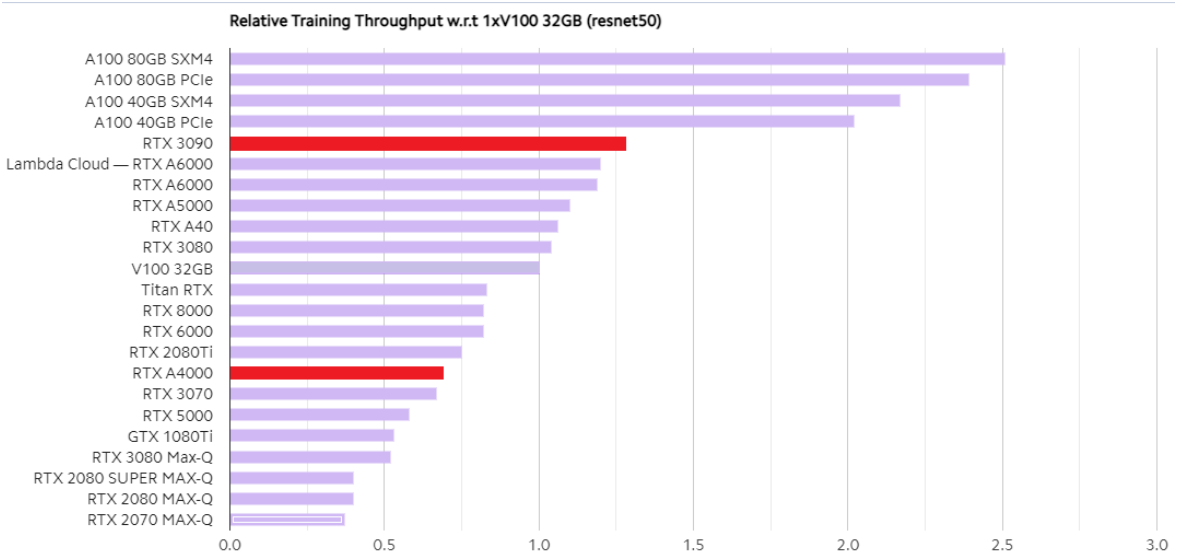
Podpora operačních systémů je velice široká. Obecným předpokladem je podpora grafických karet NVIDIA, jelikož framework využívá její CUDA jádra. Z jednotlivých linuxových distribucí jmenujme například Ubuntu verze 13.04 a novější, CentOS 7.3-1611 či Debian 8.0 a novější.

Podpora operačního systému macOS je rovněž zachována ovšem pouze pro provoz na CPU, bez možnosti použití grafické karty. Minimální verzí je macOS 10.10 (Yosemite) nebo novější.

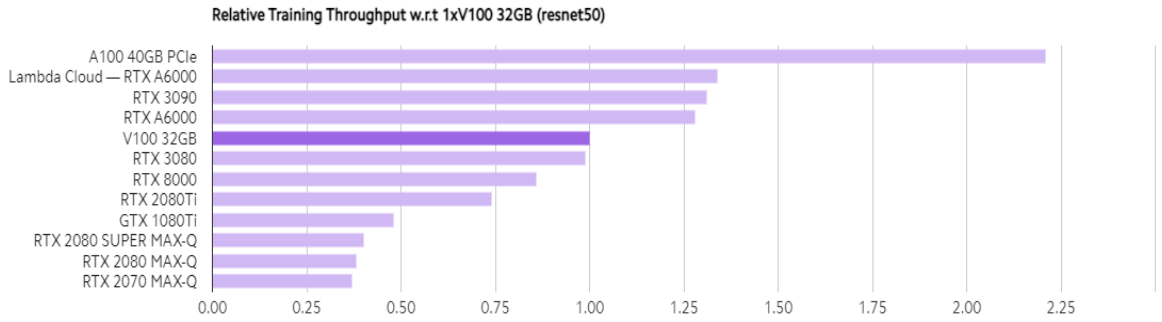
Operační systémy Windows jsou podporovány ve verzích Windows 7 a novější, ale verze Windows 10 je doporučována. Stejně tak podpora Windows Serveru 2008 a novější. [45]

5 VÝKONOVÉ SROVNÁNÍ DVOU SYSTÉMŮ

Na grafech níže si můžeme všimnout výkonového srovnání napříč grafickými kartami běžícími na systémech s frameworky TensorFlow a PyTorch. Metodou měření je tzv. tréninková propustnost, která měří počet vzorků zpracovaných grafickou kartou za sekundu. Právě propustnost velice úzce souvisí s dobou vyřešení zadaného problému, jelikož čím vyšší tréninkovou propustnost grafická karta má, tím rychleji je schopna zpracovat zadanou sadu dat a model se rychleji naučit. Samozřejmě vyššího výkonu můžeme docílit použitím většího počtu grafických karet. Použitým modelem je ResNet-50.



Obrázek 21 Srovnání napříč GPU na frameworku PyTorch [46]



Obrázek 22 Srovnání napříč GPU na frameworku TensorFlow. Zdroj:[46]

Výsledky srovnání jednotlivých grafických karet na těchto dvou systémech jsou si velice podobné. Ku příkladu vezmu GPU řady RTX 3090, která v případě PyTorch dosahuje hodnoty 1,28. Ve srovnání s TensorFlow se jedná o hodnotu 1,31, což je rozdíl v řádu do 3 %.

II. PRAKTICKÁ ČÁST

6 HARDWARE PRO VÝPOČETNÍ SERVER

V teoretické části jsem navrhl tři systémy s využitím v AI a Deep learningu. V praktické části si sestavíme výpočetní server, na kterém porovnáme výkonnost a efektivitu celkem dvou grafických karet. Instalace, měření a testování systémů bude probíhat na sestavené výpočetní stanici z vlastního pořízeného hardwaru. Jednotlivé použité komponenty ve stanice nalezneme v následující tabulce.

Tabulka 6 Seznam komponent použitých ve výpočetní sestavě

Komponenta	*Název komponenty*
Počítačová skříň	4U rack s šesticí ventilátorů
Procesor	AMD Threadripper 1920X
Chladič procesoru	Noctua NH-U9 TR4-SP3
74cryzehu Základní deska	ASRock Taichy X399
Diskové úložiště	2ks Samsung 970 EVO Plus 500 GB
Operační paměť	4ks Crucial modul DDR4 8 GB 2400Mhz s podporou ECC
Zdroj	EVGA Supernova 1300 W G2
Grafické karty	ZOTAC Gaming GeForce RTX 3090 24 GB NVIDIA A4000 16 GB

Veškeré komponenty mimo grafické karty byly pořízeny za celkovou cenu 34 030 Kč. První ze jmenovaných grafických karet značky ZOTAC byla pořízena za cenu 54 586 Kč. Druhá karta od společnosti NVIDIA byla pořízena za 32 490 Kč. Pro účely této práce tedy vznikly celkem dvě výpočetní stanice s totožným základním hardwarem, ale rozdílnými grafickými kartami.

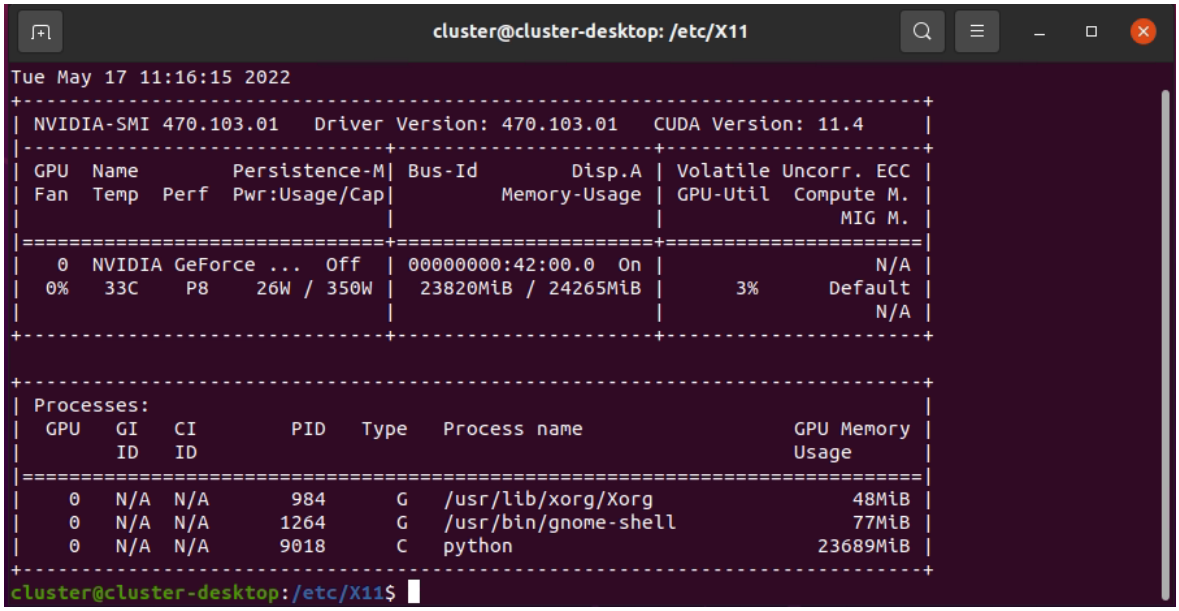
Tabulka 7 Ceny jednotlivých výpočetních serverů

Název výpočetní stanice	*Použitá grafická karta*	*Celková cena serveru*
Cluster01	ZOTAC GeForce RTX 3090 24 GB	83 616 Kč
Cluster02	NVIDIA A4000 16 GB	66 520 Kč

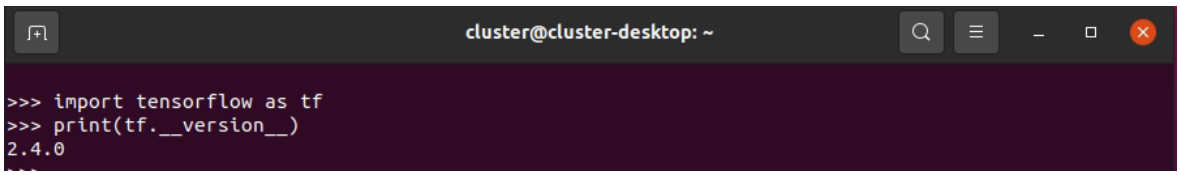
Obrázek 23 Náhled do útrob výpočetního serveru

7 OPERAČNÍ SYSTÉM A TENSORFLOW

Na výpočetní server jsem nainstaloval operační systém Ubuntu verze 20.04. Do systému bylo nutné nainstalovat grafické ovladače, které jsem zvolil ve verzi 470.103.01 dle doporučení operačního systému. Jako další krok jsem doinstaloval CUDA verze 11.0.3 spolu s knihovnami cuDNN verze 8.0.5, které jsou kompatibilní právě s instalovanou verzí CUDA. Po instalaci všech těchto komponent je možné nainstalovat Tensorflow verze 2.4.0. Pro správné fungování celého systému a jeho ovládání je potřeba nainstalovat i Python verze 3.8. [40][47][48]



Obrázek 24 Výpis dat o grafické kartě včetně verze ovladačů



Obrázek 25 Ověření verze Tensorflow a správnosti instalace

8 SOFTWARE PRO BENCHMARK A JEHO INSTALACE

AI-benchmark je otevřená volně dostupná knihovna napsaná v jazyce Python pro ověření výkonu různých zařízení určených pro práci s AI. Dokáže ověřit i výkon procesoru, ale je primárně určena k práci s grafickými kartami. Velice úzce spolupracuje s knihovnami Tensorflow a díky své jednoduchosti a přesnosti se velice hodí pro hodnocení rychlosti a trénování modelů Deep learningu.

Program provede celkem 42 testů různých algoritmů. Mezi zmíněné algoritmy patří:

1. MobileNet-V2
2. Inception-V3
3. Inception-V4
4. Inception-ResNet-V2
5. ResNet-V2-50
6. ResNet-V2-152
7. VGG-16
8. SRCNN 9-5-5
9. VGG-19
10. ResNet-SRGAN
11. ResNet-DPED
12. U-Net
13. Nvidia-SPADE
14. ICNet
15. PSPNet
16. DeepLab
17. Pixel-RNN
18. LSTM
19. GNMT [49]

8.1 Instalace AI-Benchmark

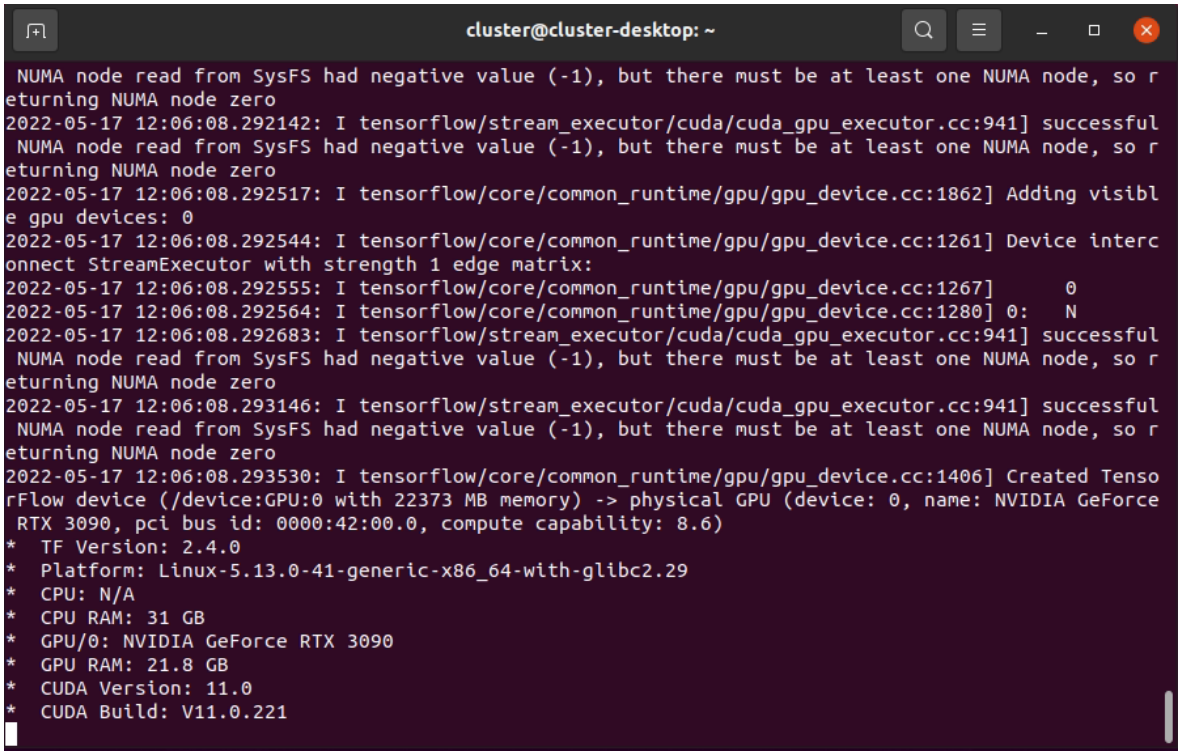
Instalaci provedeme tak, že do terminálu zadáme příkaz:

```
pip install ai-benchmark
```

Následně můžeme spouštět benchmark příkazem:

```
from ai_benchmark import AIBenchmark
benchmark = AIBenchmark()
results = benchmark.run()
```

Po spuštění benchmarku nám software v první řadě ukáže, na jakém zařízení bude provádět ověření výkonu. Následně spustí jednotlivé iterace algoritmů, kde po dokončení každého z nich nám zobrazí výsledné skóre. [50]



Obrázek 26 Ukázka spuštění softwaru pro benchmark

9 POROVNÁNÍ VÝKONNOSTI A EFEKTIVITY V JEDNOTLIVÝCH BENCHMARCÍCH

Ověření výkonu jsem provedl na celkem dvou grafických kartách, a to modelu RTX 3090 a NVIDIA A4000. Výsledek každého benchmarku jsem zapsal do tabulky včetně momentální spotřeby grafické karty. Měření spotřeby grafické karty probíhalo za pomocí wattmetru. Výsledek měření je uveden již po odečtení spotřeby ostatních komponent výpočetního serveru. Na každou grafickou kartu jsem provedl celkem tři měření a následně všechna měření zprůměroval, abych dosáhl vyšší přesnosti měřených dat. Výsledky měření byly zaokrouhleny na dvě desetinná místa.

9.1 Ověření výkonu RTX 3090

Každá jednotlivá iterace měření probíhala přibližně dvacet minut a byly u ní naměřeny hodnoty zaznamenané v tabulkách níže.

Použitý benchmark	1. kolo měření		
	Naměřené skóre	Spotřeba karty (Watt)	AI skóre
1. MobileNet-V2	162,00	117,00	40 538,00
2. Inception-V3	104,30	178,00	
3. Inception-V4	133,65	205,00	
4. Inception-ResNet-V2	141,10	255,00	
5. ResNet-V2-50	63,23	224,00	
6. ResNet-V2-152	115,30	193,00	
7. VGG-16	49,95	212,00	
8. SRCNN 9-5-5	106,00	187,00	
9. VGG-19	126,00	179,00	
10. ResNet-SRGAN	81,00	215,00	
11. ResNet-DPED	105,10	187,00	
12. U-Net	105,20	269,00	
13. Nvidia-SPADE	74,30	210,00	
14. ICNet	269,00	186,00	
15. PSPNet	177,10	267,00	
16. DeepLab	63,20	238,00	
17. Pixel-RNN	1 532,00	161,00	
18. LSTM	698,00	142,00	
19. GNMT	111,00	179,00	

Obrázek 27 První iterace měření výkonu RTX 3090

Použitý benchmark	2. kolo měření		
	Naměřené skóre	Spotřeba karty (Watt)	AI skóre
1. MobileNet-V2	160,00	116,00	40 031,00
2. Inception-V3	102,50	174,00	
3. Inception-V4	131,45	211,00	
4. Inception-ResNet-V2	140,30	248,00	
5. ResNet-V2-50	65,10	226,00	
6. ResNet-V2-152	116,60	196,00	
7. VGG-16	50,00	216,00	
8. SRCNN 9-5-5	104,00	183,00	
9. VGG-19	124,00	178,00	
10. ResNet-SRGAN	83,00	209,00	
11. ResNet-DPED	103,80	185,00	
12. U-Net	106,30	273,00	
13. Nvidia-SPADE	73,60	204,00	
14. ICNet	271,70	186,00	
15. PSPNet	179,30	269,00	
16. DeepLab	61,80	241,00	
17. Pixel-RNN	1 511,00	163,00	
18. LSTM	687,00	139,00	
19. GNMT	113,00	182,00	

Obrázek 28 Druhá iterace měření výkonu RTX 3090

V poslední tabulce, oproti předchozím, nalezneme i zmíněné průměrné výsledky všech měření. Sloupeček „Výsledné průměrné skóre“ či „Průměrné AI skóre“ značí aritmetický průměr všech bodů z předchozích tří měření. Rovněž „Průměrná spotřeba grafické karty (Watt)“ znázorňuje aritmetický průměr všech naměřených spotřeb elektrické energie ve wattech.

Použitý benchmark	3. kolo měření			Výsledek měření		
	Naměřené skóre	Spotřeba karty (Watt)	AI skóre	Výsledné průměrné skóre	Průměrná spotřeba grafické karty (Watt)	Průměrné AI skóre
1. MobileNet-V2	159,00	115,00	40 701,00	160,33	116,00	40 423,33
2. Inception-V3	103,50	176,00		103,43	176,00	
3. Inception-V4	130,90	210,00		132,00	208,67	
4. Inception-ResNet-V2	143,40	239,00		141,60	247,33	
5. ResNet-V2-50	65,28	232,00		64,54	227,33	
6. ResNet-V2-152	116,65	199,00		116,18	196,00	
7. VGG-16	51,30	219,00		50,42	215,67	
8. SRCNN 9-5-5	107,00	182,00		105,67	184,00	
9. VGG-19	124,00	178,00		124,67	178,33	
10. ResNet-SRGAN	85,00	211,00		83,00	211,67	
11. ResNet-DPED	106,70	191,00		105,20	187,67	
12. U-Net	104,85	270,00		105,45	270,67	
13. Nvidia-SPADE	75,55	211,00		74,48	208,33	
14. ICNet	270,00	185,00		270,23	185,67	
15. PSPNet	180,10	268,00		178,83	268,00	
16. DeepLab	61,40	239,00		62,13	239,33	
17. Pixel-RNN	1 523,00	164,00		1 522,00	162,67	
18. LSTM	690,00	141,00		691,67	140,67	
19. GNMT	115,00	181,00		113,00	180,67	

Obrázek 29 Poslední iterace měření spolu s průměrnými výsledky všech předchozích iterací

Karta dosahuje opravdu vysokého výkonu ve všech jednotlivých benchmarcích, ale také poměrně vysoké spotřeby elektrické energie. Se spotřebou jsou spjaté i celkové provozní náklady za výpočetní server s tímto typem grafické karty. Výsledky měření můžeme porovnat například s kartou Tesla V100 SXM2 32 GB, která v benchmarku tohoto typu dosáhla celkového AI skóre 35 791 bodů, což je o téměř 12 % méně. Jedná se už o starší hardware, ovšem v době uvedení na trh stála mnohonásobně více než testovaná RTX 3090. I nyní se dá ještě pořídit za cenu několikaset tisíc korun. [51][52]

9.2 Ověření výkonu A4000

Druhou kartou, u které jsem provedl ověření výkonu je NVIDIA A4000. Postup měření byl totožný jako u předchozího výkonnějšího modelu. V tabulkách se můžeme podívat na naměřené hodnoty.

Použitý benchmark	1. kolo měření		
	Naměřené skóre	Spotřeba karty (Watt)	AI skóre
1. MobileNet-V2	39,70	78,00	26 535,00
2. Inception-V3	58,40	92,00	
3. Inception-V4	53,80	99,00	
4. Inception-ResNet-V2	73,90	115,00	
5. ResNet-V2-50	42,10	123,00	
6. ResNet-V2-152	54,60	112,00	
7. VGG-16	61,30	120,00	
8. SRCNN 9-5-5	76,70	128,00	
9. VGG-19	74,50	89,00	
10. ResNet-SRGAN	103,00	96,00	
11. ResNet-DPED	114,00	95,00	
12. U-Net	184,00	110,00	
13. Nvidia-SPADE	68,00	121,00	
14. ICNet	153,00	130,00	
15. PSPNet	290,00	122,00	
16. DeepLab	111,00	128,00	
17. Pixel-RNN	479,00	73,00	
18. LSTM	369,00	90,00	
19. GNMT	150,00	114,00	

Obrázek 30 První iterace měření výkonu NVIDIA A4000

Použitý benchmark	2. kolo měření		
	Naměřené skóre	Spotřeba karty (Watt)	AI skóre
1. MobileNet-V2	44,70	76,00	26 123,00
2. Inception-V3	64,50	90,00	
3. Inception-V4	54,00	98,00	
4. Inception-ResNet-V2	77,80	113,00	
5. ResNet-V2-50	41,90	118,00	
6. ResNet-V2-152	58,50	115,00	
7. VGG-16	62,40	122,00	
8. SRCNN 9-5-5	76,90	127,00	
9. VGG-19	78,10	94,00	
10. ResNet-SRGAN	99,40	93,00	
11. ResNet-DPED	117,00	95,00	
12. U-Net	183,00	116,00	
13. Nvidia-SPADE	69,50	122,00	
14. ICNet	160,00	136,00	
15. PSPNet	295,00	118,00	
16. DeepLab	105,00	124,00	
17. Pixel-RNN	485,00	75,00	
18. LSTM	382,00	97,00	
19. GNMT	145,00	111,00	

Obrázek 31 Druhá iterace měření výkonu NVIIDA A4000

Použitý benchmark	3. kolo měření			Výsledek měření		
	Naměřené skóre	Spotřeba karty (Watt)	AI skóre	Výsledné průměrné skóre	Průměrná spotřeba grafické karty (Watt)	Průměrné AI skóre
1. MobileNet-V2	42,80	76,00	26 348,00	42,40	76,67	26 335,33
2. Inception-V3	62,30	93,00		61,73	91,67	
3. Inception-V4	53,90	99,00		53,90	98,67	
4. Inception-ResNet-V2	75,90	116,00		75,87	114,67	
5. ResNet-V2-50	41,30	114,00		41,77	118,33	
6. ResNet-V2-152	55,40	118,00		56,17	115,00	
7. VGG-16	63,10	121,00		62,27	121,00	
8. SRCNN 9-5-5	75,90	125,00		76,50	126,67	
9. VGG-19	74,10	87,00		75,57	90,00	
10. ResNet-SRGAN	101,50	94,00		101,30	94,33	
11. ResNet-DPED	116,00	98,00		115,67	96,00	
12. U-Net	183,00	114,00		183,33	113,33	
13. Nvidia-SPADE	69,60	124,00		69,03	122,33	
14. ICNet	154,40	132,00		155,80	132,67	
15. PSPNet	296,70	117,00		293,90	119,00	
16. DeepLab	108,80	126,00		108,27	126,00	
17. Pixel-RNN	478,00	74,00		480,67	74,00	
18. LSTM	374,00	97,00		375,00	94,67	
19. GNMT	149,00	113,00		148,00	112,67	

Obrázek 32 Poslední iterace měření výkonu NVIDIA A4000

9.2.1 Závěr z testování

Testovaná grafická karta NVIDIA A4000 je oproti první z testovaných karet RTX 3090 téměř o 35 % pozadu v AI skóre. Co je ale důležité zmínit, že A4000 sice dosáhla menšího celkového skóre, ale zároveň je o více jak 53 % úspornější na spotřebě elektrické energie. Průměrná spotřeba RTX 3090 za všechny testované benchmarky je 200 wattů, zatím co u A4000 je to výrazně méně, rovných 107 wattů. Zároveň disponuje pamětí typu ECC, tudíž může být pro provoz v serveru vhodnější.

9.3 Srovnání efektivity při zadání stejného výpočtu

V následující kapitole provedu porovnání efektivity navržených grafických karet na základě výsledků z jejich benchmarků. Výpočet je orientační a nejsou v něm zahrnuty odchylky v účinnosti zdroje. K jednotlivým výpočtů použiji následující parametry:

- Průměrné AI skóre z benchmarků.

- Výsledná průměrná spotřeba jednotlivé grafické karty.
- Předpokládejme řešenou úlohu, která bude RTX 3090 trvat 14 dnů, v přepočtu 336 hodin.
- Cena za spotřebovanou kilowatthodinu (kWh) je rovna 2 Kč.

9.3.1 Doba řešení úlohy na A4000

Nejprve je nutné zjistit, o kolik hodin déle oproti RTX 3090 bude úlohu řešit karta A4000. Z důvodu jednoduššího výpočtu jsem výsledné průměrné skóre z každé grafické karty zaokrouhlil na celé číslo. Ze zadaných parametrů víme, že RTX 3090 řeší úlohu 14 dnů a průměrné skóre v AI benchmarku bylo 40 423 bodů. Karta A4000 dosáhla průměrného skóre 26 335 bodů.

$$rozdíl\ výkonu = \frac{26335}{40423} * 100 = 65{,}14\%$$

$$rozdíl\ výkonu = 100 - 65{,}14 = 34{,}86\% \approx \mathbf{35}\%$$

Z výše uvedené rovnice vyplývá, že grafická karta A4000 je o 35% pomalejší, než RTX 3090. Nyní tedy můžeme odhadnout, jak dlouho by asi výpočet trval na A4000.

$$doba\ výpočtu\ na\ A4000 = 14 * 0{,}35 = 4{,}9 + 14 = \mathbf{18{,}9}$$

Výsledkem je, že v případě spuštění totožné úlohy na kartě A4000 se dá předpokládat odhadovaná doba vyřešení za 18,9 dne, v přepočtu 453,6 hodin.

9.3.2 Spotřebovaná elektřina

Dále je nutné stanovit, kolik za svůj provoz spotřebuje elektřiny každá z grafických karet. Známe vstupní parametry, tedy počet hodin řešení úlohy, průměrnou spotřebu a cenu za kilowatthodinu. Výpočtem dostaneme spotřebu udávanou ve wattech. Proto celou rovnici vydělíme tisícem z důvodu převodu na kilowatthodiny.

$$Spotřebovaná\ el. = \frac{průměr\ spotřeby * počet\ hodin\ v\ provozu}{1000}$$

$$Spotřebovaná\ elektřina\ RTX3090 = \frac{200 * 336}{1000} = \mathbf{67{,}2\ kWh}$$

$$Spotřebovaná\ elektřina\ A4000 = \frac{107 * 453{,}6}{1000} = \mathbf{48{,}54\ kWh}$$

Výsledkem výpočtu je, že grafická karta RTX 3090 spotřebuje za řešenou úlohu 67,2 kWh. Spotřeba grafické karty A4000 je 48,54 kWh.

9.3.3 Cena spotřebované elektřiny

Výsledky z předchozího výpočtu použijeme ke stanovení ceny elektřiny.

$$Cena\ spotřebované\ elektřiny = spotřebované\ kWh * cena\ za\ kWh$$

$$Cena\ za\ spotřebovanou\ elektřinu\ RTX\ 3090 = 67{,}2 * 2 = \mathbf{134{,}4\ korun}$$

$$Cena\ za\ spotřebovanou\ elektřinu\ A4000 = 48{,}54 * 2 = \mathbf{97{,}07\ korun}$$

Kombinací předchozích výpočtů jsem dospěl k závěru, že provoz úlohy, která by trvala 14 dnů na grafické kartě RTX 3090 by stála 134,4 Kč na spotřebované elektřině. Druhé z karet by ta stejná úloha trvala nejspíše 18,9 dnů a spotřebovala by elektřinu v hodnotě 97,07 Kč.

9.3.4 Závěr

Závěrem bych zmínil modelovou úlohu, která by grafické kartě RTX 3090 trvala přesně 365 dní, tedy 1 rok. Po dosazení do předchozích vzorců zjistíme, že za 1 rok řešení úlohy spotřebuje 1 752 kWh elektřiny v celkové hodnotě 3 504 korun. Na tomto příkladu už se projeví menší výkon karty A4000, které by stejná úloha trvala o více jak 492 dnů déle, ovšem spotřebovala by o 486 kWh méně elektřiny v celkového hodnotě 2 530,76 Kč. Vzhledem k tak velkému nárůstu času je na zvážení, zda by se koupě méně výkonné karty vyplatila. Záleží na posouzení uživatele, zda je ochoten přijmout nárůst řešení úlohy v čase s nižší spotřebou a pořizovací cenou, nebo zaplatit vyšší cenu za pořízení, snášet celkově vyšší provozní náklady spojené s elektřinou, ovšem s kratší dobou řešení zadané úlohy.

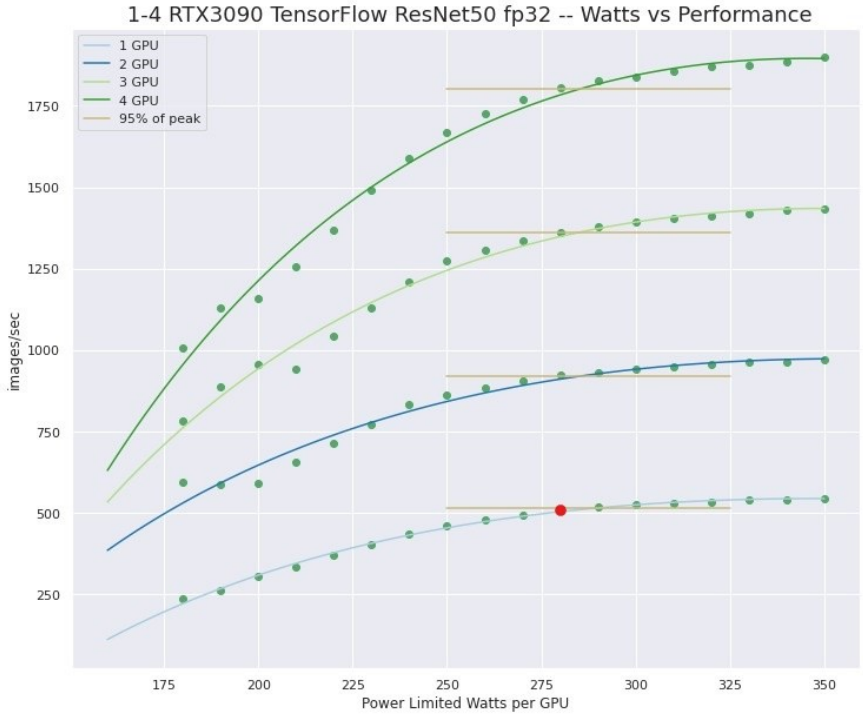
10 MOŽNOSTI OPTIMALIZACE ZDROJŮ

V době zdražující elektřiny je potřeba se zamýšlet i nad způsoby, jakými bychom mohli uspořit část nákladů vydaných na provoz výpočetního serveru. Existují způsoby, jakými můžeme snížit spotřebovanou energií či zefektivnit celý proces.

10.1 Snížení maximální spotřeby grafické karty

Každá grafická karta má maximální spotřebu, které může dosáhnout v maximální zátěži. V teoretické částí jsem ji popisoval jako TDP. Jak již bylo zmíněno, karta RTX 3090 dosahuje hodnoty TDP 350 wattů. Optimalizací můžeme tedy dosáhnout toho, že snížíme tuto maximální hranici, čímž dojde ke snížení spotřeby. Rovněž můžeme se spotřebou různě pracovat a najít optimální poměr mezi spotřebovanou elektrickou energií a poskytnutým výkonem.

Na grafu níže můžeme pozorovat provoz serveru s 1–4 grafickými kartami řady RTX 3090. Žlutá přímka na grafu znázorňuje hranici devadesáti pěti procentního výkonu grafické karty. Světle modrá křivka ve spodní části grafu nám ukazuje, jak výkon klesá v okamžiku, kdy se blíží maximální možné spotřebě grafické karty. Z grafu je tedy patrné, že optimálním poměrem mezi spotřebou elektrické energie a téměř maximálním výkonem je nastavení limitu TDP na cca 280 wattů. Použitým benchmarkem byl již zmiňovaný ResNet-50. [53]



Obrázek 33 Porovnání výkonu po úpravě spotřeby grafické karty [53]

10.2 Optimální vytížení grafických karet

Další možností jak optimalizovat náklady na spotřebovanou elektřinu je efektivně nakládat s výpočetním serverem a jeho provozem. Optimálním řešením by bylo, aby byl server pokud možno konstantně vytížen a neměl žádná prázdná místa, kdy by běžel bez zatížení, bez jakékoliv práce. Toho můžeme docílit použitím aplikace zvané Tensorflow Profiler, která dokáže monitorovat používané modely Tensorflow a jejich komunikace mezi procesorem či grafickou kartou. Umožňuje tedy uživateli zobrazit aktuálně využívané hardwarové prostředky a Tensorflow operace, u nichž můžeme najít nějaké abnormality.

Obrázek 34 Mezery mezi jednotlivými kroky značí přerušení vytížení grafické karty [54]

Na grafu výše si můžeme všimnout, že mezi operací číslo osm a devět je poměrně velká časová prodleva. Jinak řečeno, v tuto chvíli naše grafická karta byla po nějaký čas nevyužitá a spotřebovávala elektrickou energie zbytečně. Analýzou a optimalizací všech těchto prázdných kroků můžeme docílit optimálního a konstantního vytížení grafické karty a minimalizovat zbytečné prostoje. [54][55]

11 ZABEZPEČENÍ SYSTÉMU PŘED NEOPRÁVNĚNÝMI PŘÍSTUPY

Každý server je zapotřebí zabezpečit před přístupem neoprávněných osob. U osobních počítačů se často jedná o nasazení antivirového programu. V případě serverů máme různé možnosti, jakým způsobem můžeme server zabezpečit.

11.1 Pravidelné aktualizace systému

Každý administrátor jakéhokoliv serveru by měl udržovat veškeré systémy, pokud možno aktuální. Jedná se především o tzv. security updates, kde se vývojáři daného systému zaměřují na opravu chyb, které objeví v průběhu provozu systému. Může se stát, že v zastaralé verzi systému bude objevena nějaká závažná bezpečnostní chyba, kterou vývojáři opravili v následující verzi, ovšem administrátor neprovedl její aktualizaci. Útočník tak může zneužít této zranitelnosti a využít ji ve svůj prospěch.

11.1.1 Konfigurace automatických aktualizací

Nejprve je nutné do systému doinstalovat baličky, které umožňují bezobslužnou instalaci aktualizací. Po otevření příkazové řádky tedy zadáme příkazy:

```
sudo apt install unattended-upgrades
```

```
sudo apt install update-notifier-common
```

Následně je zapotřebí provést konfiguraci. Pokračujeme tedy příkazem:

```
sudo nano /etc/apt/apt.conf.d/50unattended-upgrades
```

Otevře se nám konfigurační soubor, ve kterém můžeme nakonfigurovat způsob, jakým se budou aktualizace instalovat. Především se jedná o konfiguraci:

- Možnost zadat email, na který přijdou notifikace v případě výskytu nějakého problému.
- Nastavení automatického odstranění staré aktualizace či nepoužívaných balíčků.
- Automatický restart systému po aktualizaci, případně nastavení času restartu.

V našem případě jsem nastavil, aby se server restartoval vždy ve dvě hodiny ráno a v případě problému odeslání e-mailu na příslušnou emailovou adresu.

```
// Install all updates when the machine is shutting down
// instead of doing it in the background while the machine is running.
// This will (obviously) make shutdown slower.
// Unattended-upgrades increases logind's InhibitDelayMaxSec to 30s.
// This allows more time for unattended-upgrades to shut down gracefully
// or even install a few packages in InstallOnShutdown mode, but is still a
// big step back from the 30 minutes allowed for InstallOnShutdown previously.
// Users enabling InstallOnShutdown mode are advised to increase
// InhibitDelayMaxSec even further, possibly to 30 minutes.
//Unattended-Upgrade::InstallOnShutdown "false";

// Send email to this address for problems or packages upgrades
// If empty or unset then no email is sent, make sure that you
// have a working mail setup on your system. A package that provides
// 'mailx' must be installed. E.g. "user@example.com"
//Unattended-Upgrade::Mail "report@seznam.cz";

// Set this value to one of:
//    "always", "only-on-error" or "on-change"
// If this is not set, then any legacy MailOnlyOnError (boolean) value
// is used to chose between "only-on-error" and "on-change"
//Unattended-Upgrade::MailReport "on-change";

// Remove unused automatically installed kernel-related packages
// (kernel images, kernel headers and kernel version locked tools).
//Unattended-Upgrade::Remove-Unused-Kernel-Packages "true";

// Do automatic removal of newly unused dependencies after the upgrade
//Unattended-Upgrade::Remove-New-Unused-Dependencies "true";

// Do automatic removal of unused packages after the upgrade
// (equivalent to apt-get autoremove)
//Unattended-Upgrade::Remove-Unused-Dependencies "false";

// Automatically reboot *WITHOUT CONFIRMATION* if
//  the file /var/run/reboot-required is found after the upgrade
//Unattended-Upgrade::Automatic-Reboot "false";

// Automatically reboot even if there are users currently logged in
// when Unattended-Upgrade::Automatic-Reboot is set to true
//Unattended-Upgrade::Automatic-Reboot-WithUsers "true";

// If automatic reboot is enabled and needed, reboot at the specific
// time instead of immediately
//  Default: "now"
//Unattended-Upgrade::Automatic-Reboot-Time "02:00";
```

Obrázek 35 Ukázka konfiguračního souboru 50unattended-upgrades

Po nastavení všech parametrů v konfiguračním souboru pokračujeme příkazem:

dpkg-reconfigure -plow unattended-upgrades

Po zadání příkazu se nám zobrazí fialová tabulka, kde stojí, zda chceme opravdu automaticky stahovat a instalovat stabilní aktualizace. Tabulku samozřejmě odsouhlasíme. Po

jejím zmizení se nám ve složce /etc/apt/apt.conf.d/ vytvořil soubor 20auto-upgrades. Soubor obsahuje celkem dva řádky s číslicí na konci. To nám říká, že konfigurace byla správná a aktualizace se budou instalovat automaticky. [56]

Obrázek 36 Obsah konfiguračního souboru 20auto-upgrades

11.2 Vytvoření uživatele mimo hlavního administrátora

V linux systémech je absolutně nejvyšším uživatelským oprávněním tzv. root. Tento uživatel může provádět veškeré systémové operace, instalace, spouštět kódy atd... Proto je nežádoucí, aby administrátor či dokonce uživatel neustále pracoval s touto úrovní oprávnění. V případě potřeby může zadáním příkazu sudo a příslušného hesla dočasně zvýšit svá oprávnění.

Tímto jednoduchým příkazem dokážeme zjistit, kteří uživatelé jsou zařazeni do skupiny root a mají tak plný přístup k systému:

```
grep -Po '^sudo.+:\K.*$' /etc/group
```

11.3 Složitost hesla

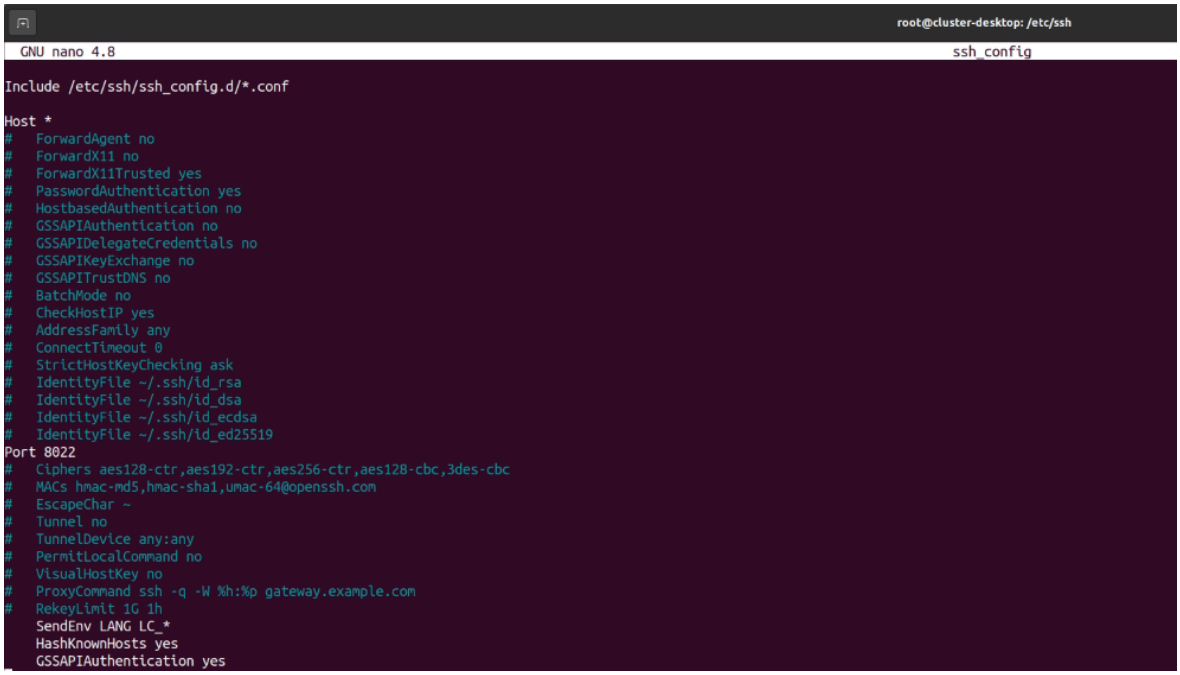
Heslo by mělo být složeno z běžných znaků, číslic a speciálních znaků v patřičné délce. Spousta uživatelů ovšem volí velice krátká hesla, která nesplňují bezpečnostní požadavky a útočník tak může relativně jednoduše takové heslo prolomit. Příkazem nano /etc/security/pwquality.conf se dostaneme do konfiguračního souboru, ve kterém můžeme nastavit minimální délku hesla, počet požadovaných velkých, malých či speciálních znaků.

```
# Configuration for systemwide password quality limits
# Defaults:
#
# Number of characters in the new password that must not be present in the
# old password.
# difok = 1
#
# Minimum acceptable size for the new password (plus one if
# credits are not disabled which is the default). (See pam_cracklib manual.)
# Cannot be set to lower value than 6.
# minlen = 8
#
# The maximum credit for having digits in the new password. If less than 0
# it is the minimum number of digits in the new password.
# dcredit = 0
#
# The maximum credit for having uppercase characters in the new password.
# If less than 0 it is the minimum number of uppercase characters in the new
# password.
# ucredit = 0
#
# The maximum credit for having lowercase characters in the new password.
# If less than 0 it is the minimum number of lowercase characters in the new
# password.
# lcredit = 0
#
# The maximum credit for having other characters in the new password.
# If less than 0 it is the minimum number of other characters in the new
# password.
# ocredit = 0
#
# The minimum number of required classes of characters for the new
# password (digits, uppercase, lowercase, others).
# minclass = 0
```

Obrázek 37 Konfigurační kód pro zabezpečení systémů heslem

11.4 Změna portu Security shell (SSH)

Aby bylo možné se na systémy vzdáleně připojovat, každý systém poskytuje SSH server na síťovém portu 22. Problém je, že tohoto jsou si vědomi i potencionální útočníci, kteří mohou různými technikami zkoušet odhadnout jméno a heslo uživatele a dostat se tak do systému. Internetem navíc koluje spousta automatických robotů, kteří prozkoumávají různé IP adresy a hledají, zda náhodou není na výchozí bráně, tedy routeru, povolený přístup právě na port 22. Tuto situaci můžeme vyřešit změnou portu v systému na jiný, námi zvolený port, který ovšem nemáme obsazený jinou aplikací. Otevřeme si tedy konfigurační soubor za pomocí příkazu sudo nano /etc/ssh/ssh_config. Zde vyhledáme řádek, který obsahuje část Port 22. Z řádku odstraníme počáteční znak mřížku a port změníme na námi zvolený. Soubor po naší úpravě tedy může vypadat následovně.



Obrázek 38 Změna SSH portu z výchozího na vlastní

Nakonec provedeme restart služby, která se stará o provoz SSH serveru, a to příkazem sudo service ssh restart.

11.5 Vyžadování bezpečnostního klíče pro přihlášení

Dalším krokem, kterým můžeme zvýšit úroveň zabezpečení, je změnit výchozí nastavení způsobu připojení k serveru pomocí Security shell. Ve výchozím nastavení se připojujeme pomocí jména a námi zvoleného hesla. Systém ovšem umožňuje změnit způsob přihlašování za pomocí 3072bitového RSA[4] klíče, pomocí kterého se následně může administrátor autentizovat bez nutnosti zadávat heslo. Jedná se o pár veřejného a privátního klíče, kde veřejný klíč je obsažen na cílovém serveru a soukromý klíč na straně klienta. Pokud se klient chce připojit ke vzdálenému serveru, server zašifruje náhodný soubor znaků veřejným klíčem, který je schopen dešifrovat pouze ten, kdo má soukromý klíč v páru s veřejným klíčem. Jde o bezpečnější metodu přihlašování do systému oproti klasickému způsobu s heslem. Standartně je klíč generován o délce 3072 bitů. Je možné použít parametr, který může vygenerovat klíč o délce až 4096 bitů a zvýšit tak ještě úroveň zabezpečení.

[4] RSA – algoritmus sloužící mimo jiné k šifrování komunikace

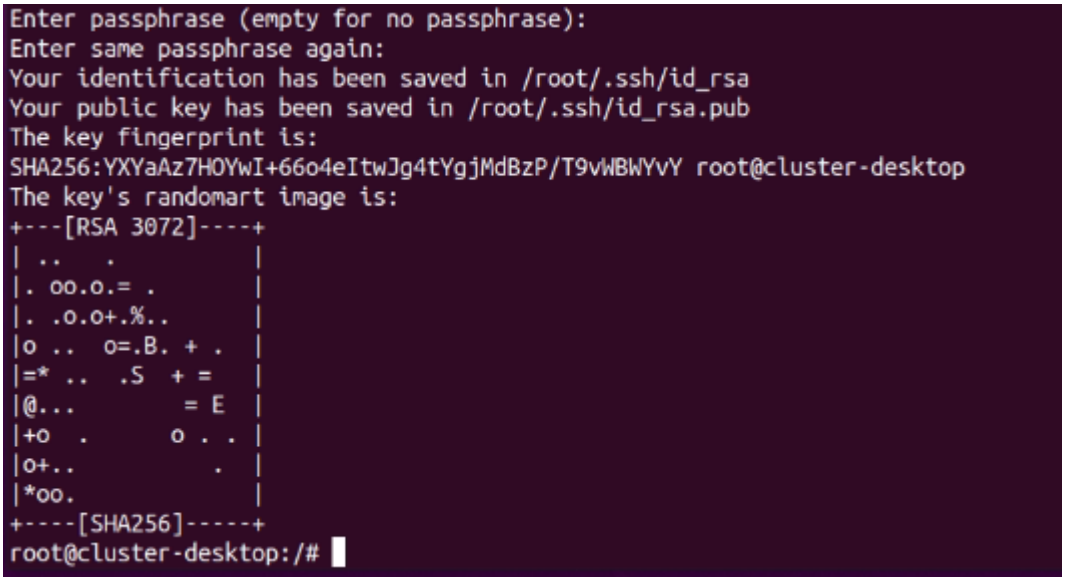
11.5.1 Způsob generování klíčů

Nejprve provedeme vygenerování páru klíčů na klientském počítači pomocí příkazu:

ssh-keygen

Tímto příkazem spustíme proces generování klíče. Pokud bychom chtěli generovat klíč o větší délce, stačí za zmiňovaný příkaz přidat parametr -p 4096.

Následně se nás systém dotáže, kam chceme výsledný generovaný klíč uložit. Výchozí nastavení můžeme změnit zadáním vlastní cesty k souboru. Dalším krokem je možné zadat passphrase. Pro potřeby této práce si proto zvolím slovo o délce 15 znaků s kombinací malých, velkých písmen, číslic i speciálních znaků. O úspěšném vygenerování páru klíčů by nás měl systém informovat výstupem, který můžeme vidět na obrázku.



Obrázek 39 Text značí, že došlo k úspěšnému vygenerování klíčů

Nakonec musíme přenést vygenerovaný veřejný klíč na cílový server, na který se chceme přihlašovat. Nejjednodušší způsob je použít příkaz:

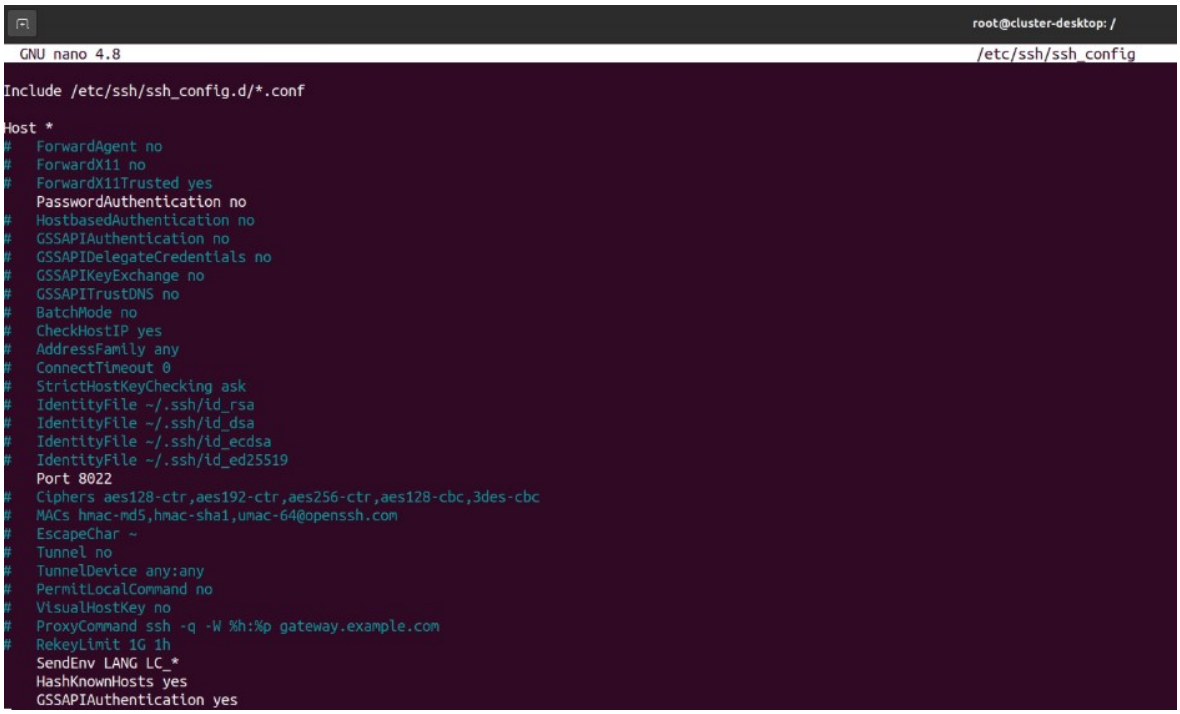
ssh-copy-id nazev_serveru@nazev_hostitele

V našem případě tedy příkaz vypadal následovně:

ssh-copy-id cluster@192.168.1.50

Tímto příkazem dojde k připojení na vzdálený server, který nás vyzve k zadání původního přihlašovacího hesla, které jsme zadávali v kombinaci s uživatelským jménem. V posledním kroku již pouze deaktivujeme způsob přihlašování pomocí jména a hesla a provedeme restart služby, která se stará o provoz SSH komunikace. Zadáme příkaz sudo nano

/etc/ssh/ssh_config a vyhledáme řádek PasswordAuthentication. Odstraníme ze začátku řádku znak mřížky a změníme konec z „yes“ na „no“.



Obrázek 40 Deaktivace přístupu k serveru heslem

Nakonec spustíme příkaz sudo service ssh restart, čímž dojde k restartování služby. Nyní se do cílového serveru dokážeme připojit pouze my, jelikož právě naše klientská stanice obsahuje privátní klíč. [58][59]

11.6 Zablokování přístupu po neúspěšných pokusech o přihlášení

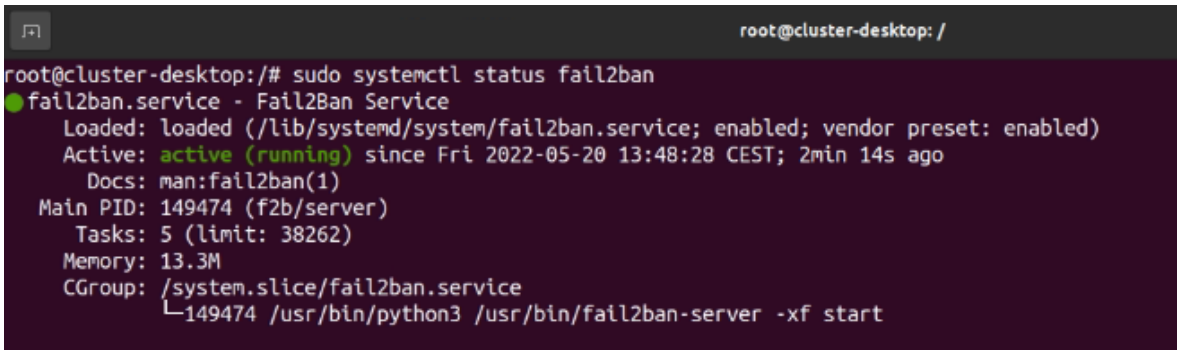
Útočníci mohou napadat server i útoky hrubou silou, tedy snažit se nabourat tak, že budou neustále zkoušet uhodnout přístupové heslo do serveru. Ve výchozím nastavení není žádná ochrana před tímto typem útoku. Existuje ovšem služba s názvem Fail2ban, která prochází autentizační log serveru a monitoruje všechny neúspěšné pokusy o přihlášení. Pokud se útočník snaží prolomit heslo, může tato služba po počtu námi nastavených neúspěšných pokusech a s pomocí systémového firewallu zakázat příslušnou internetovou adresu, odkud se útočník připojuje na určitou dobu. Po uplynutí stanovené doby je komunikace opět možná.

11.6.1 Instalace služby Fail2ban

Instalace je velice jednoduchá. Do terminálu na serveru zadáme příkaz:

sudo apt install fail2ban

Po dokončení příkazu se můžeme ujistit, že instalace proběhla správně, a to příkazem sudo systemctl status fail2ban. Měl by se nám zobrazit text, který nás informuje o běžící službě.



Obrázek 41 Informace o běžící službě Fail2ban

11.6.2 Nastavení

Konfiguraci provádíme v souboru /etc/fail2ban/jail.conf. Můžeme nastavit různé parametry. Mezi ty základní řadíme:

- maxretry – maximální počet neúspěšných pokusů o připojení,
- bantime – čas, po který bude zdrojová internetová adresa zablokována po určeném počtu neúspěšných pokusů,
- ignoreip – zadané internetové adresy nebudou nikdy blokovány,
- findtime – doba, za kterou musí dojít k dosažení maximálního počtu neúspěšných pokusů, aby došlo k blokování útočníka.

V mém nastavení jsem zvolil, aby po maximálně třech neúspěšných pokusech po dobu deseti minut byl útočník zablokován na dobu celkem sedmi dnů. Zároveň jsem zadal internetovou adresu své klientské stanice, abych ji vyloučil ze seznamu monitorovaných adres. [60]

```
# "ignoreip" can be a list of IP addresses, CIDR masks or DNS hosts. Fail2ban
# will not ban a host which matches an address in this list. Several addresses
# can be defined using space (and/or comma) separator.
ignoreip = 192.168.1.51/8 ::1

# External command that will take an tagged arguments to ignore, e.g. <ip>,
# and return true if the IP is to be ignored. False otherwise.
#
# ignorecommand = /path/to/command <ip>
#ignorecommand =

# "bantime" is the number of seconds that a host is banned.
bantime  = 7d

# A host is banned if it has generated "maxretry" during the last "findtime"
# seconds.
findtime  = 10m

# "maxretry" is the number of failures before a host get banned.
maxretry = 3
```

Obrázek 42 Konfigurace služby Fail2ban

Slabinou tohoto nástroje může být to, že útočník může využívat různé způsoby, které mu umožní měnit svou internetovou adresu a teoreticky by měl být po každém zablokování schopen si internetovou adresu změnit a pokračovat dále v útocích.

11.7 Nástroje AlienVault

Společnost AT&T Cybersecurity je americká společnost, která vyvíjí služby a aplikace, díky kterým je možné monitorovat a řešit kybernetické útoky či narušení systémů. Zabývá se tedy nástroji pro pokročilou detekci bezpečnostních hrozeb a zjištěných zranitelností v systémech. Jedním z takových nástrojů je i OTX (Open Threat Exchange). Síla tohoto nástroje spočívá v jeho otevřené komunitě, která čítá přes 130 000 členů z více jak 140 zemí z celého světa, kteří denně přispějí více jak 20 miliony nalezených hrozeb a zranitelností. Nástroj je dostupný pro kohokoliv a může být integrován do jakéhokoliv serveru či systému v cloudu, jako je Google Cloud Platform, Azure či Amazon Web Services. Umožňuje napříč celé komunitě spolupracovat na hrozbách a navzájem se informovat o nalezených zranitelnostech.

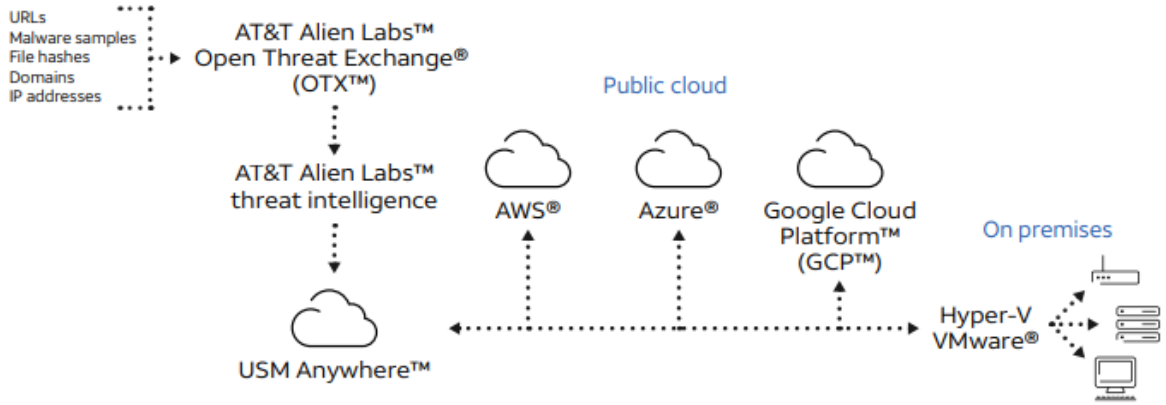
11.7.1 AlienVault USM Anywhere

USM Anywhere detekuje hrozby, které nastanou v monitorovaném systému či síti a informuje administrátora na jedné platformě. Je poskytována formou cloudové služby a umožňuje správci:

- Monitorovat prostředky.
- Vyhodnocovat zranitelnosti v reálném čase.

- Detekovat narušení.
- Behaviorální monitoring.
- Správa SIEM (Security information and event management) protokolů.

Velice úzce aplikace spolupracuje i s nástrojem OTX, ze které čerpá veškeré hrozby nahlášené právě komunitou. Obrázek níže velice dobře popisuje, jakým způsobem aplikace funguje. Veškerá vstupní data pochází z webových adres, domén, IP adres atd…Následně prochází přes OTX, kde jsou zaznamenány veškeré zranitelnosti a hrozby, přes threat intelligence nástroj, který hrozby vyhodnotí do aplikace USM Anywhere, která spojuje naše prostředky s informacemi od AlienLabs.



Obrázek 43 Princip detekce hrozeb a zranitelností USM Anywhere [62]

Jedná se ovšem o placený nástroj, kde cena začíná v době psaní této práce (květen 2022) na 1070$ a končí na 2595$. [61][62][63][64]

ZÁVĚR

Specifikací možností výběru hardwaru jsem dospěl k závěru, že na trhu existuje velké množství specializovaných výpočetních zařízení, které jsou svými výrobci konstruovány právě pro potřeby svých zákazníků, kteří si mohou v závislosti na svých požadavcích na výkon, dle svých finančních možností a účelu použití vybírat. Zmínil jsem celkem dvě varianty profesionálních serverů na trhu, které jsou poměrně dost výkonné ke zvládání těch nejnáročnějších požadavků, nicméně většinou se jedná o dost drahá řešení. Proto jsem provedl návrh vlastního řešení, u kterého se domnívám, že může splňovat základní požadavky subjektu, který se chce zabývat oblastí AI a Deep learningu.

Možností jak monitorovat tato zařízení je velké množství. Ať už přímo využití nástrojů výrobců, tak implementací vlastních řešení. Výrobci vyvíjí monitorovací software, který je součástí jejich zařízení. V případě realizace vlastních návrhů jsou na trhu volně dostupné aplikace, které mohou být implementovány, jako je například zmiňovaný Zabbix, který je navržen velice komplexně a poskytuje tak relevantní a užitečná data. Veškerá data sbírá v reálném čase a je jen a pouze na administrátorovi systému, jakým způsobem bude s daty nakládáno. Nicméně nic nebrání administrátorovi využít kombinaci softwaru dodávaného výrobcem, ale i jiných nástrojů, jako je právě Zabbix.

Dalším úkolem bylo navržení tří systémů, které je možné využít v oblasti AI a Deep learningu. Navrhnul jsem tedy tři systémy. Prvním z nich byl Keras, dále PyTorch a Tensorflow. Na systému Tensorflow jsem provedl i srovnání výkonu navrhovaných grafických karet RTX 3090 a NVIDIA A4000. Dospěl jsem k výsledku, že první ze zmiňovaných karet je vysoce výkonná a podle toho i dosahuje vysokých hodnocení ve výkonových testech. Ovšem vysoký výkon není zadarmo a je vykoupen poměrně vysokou spotřebou elektrické energie, která bude stoupat v případě zapojení i dalších těchto i dalších karet z důvodu škálování výkonu. V důsledku stoupání spotřeby se též zvyšují náklady spojené s provozem takového serveru a ty by neměly být zanedbány. Provedl jsem tak základní porovnání efektivity navržených karet, kde jsem dospěl k výsledku, že karta A4000 lepší efektivitu v závislosti na spotřebě elektrické energie než RTX 3090. Může se jednat o rozhodující faktor v případě volby, kterou ze zmiňovaných karet použít. Ovšem pokud uživatele systému nezajímá spotřeba, ale pouze co nejvyšší možný výkon, volba připadá na RTX 3090.Rovněž jsem dospěl k závěru, že výpočty na kartě A4000 jsou pomalejší oproti RTX 3090, jelikož se jedná o méně výkonnou kartu.

Mezi možnosti optimalizace zdrojů jsem zařadil snížení spotřeby elektrické energie a optimalizace úloh na grafické kartě. Díky vývojářům ovladačů je možné snižovat maximální hodnotu spotřeby grafické karty, čímž sice dochází ke snížení výkonu, ale zároveň i ke snížení spotřeby. Tento poměr může být optimalizován právě tak, aby došlo k co nejvyššímu snížení spotřeby, a zároveň co nejnižšímu snížení výkonu. Z dostupných zdrojů jsem dospěl k výsledku, že konkrétně u RTX 3090 může dojít právě ke snížení spotřeby z původní maximální hranice 350 wattů na 280 wattů. Docílíme tak snížení spotřeby o 20 %, přičemž dojde pouze k 5% snížení výkonu. Dalším zmiňovaným nástrojem je optimalizace běhu úloh v prostředí Tensorflow. Touto optimalizací nedocílíme snížení spotřeby, ale jejímu optimálnímu a kontinuálnímu využití.

V rámci provedení zabezpečení systému před neoprávněným vniknutím cizích osob či pokusu útočníků infikovat server škodlivým softwarem jsem dospěl jsem k závěru, že jedním ze základních úkonů, který musí každý z administrátorů provádět, je pravidelně aktualizovat operační systém, jelikož právě v něm může být odhalena nějaká trhlina, kterou může útočník využít k infikaci, a která bývá zpravidla odstraněna v systémových aktualizacích. Dále je poměrně spolehlivým zabezpečením změna způsobu autentifikace do systému, kde administrátor nebude využívat jméno a heslo, ale zvolí autentifikaci pomocí privátního klíče. Zabezpečení může být zvýšeno i změnou výchozího přístupového portu, který je stejný pro každou výchozí instalaci operačního systému a je známý i potencionálním útočníkům. V případě útoku hrubou silou, tedy násilnému pokusu útočníka prolomit heslo pomocí hledání jeho různých variant, je spolehlivým nástrojem i zmíněný Fail2ban, který umožní útočníka odpojit a na zvolenou dobu zakázat komunikaci s ním. Slabinu tohoto řešení vidím v tom, že útočník může využívat různých služeb, které mu umožní změnit si svou internetovou adresu a provádět tak dále útoky na server. Proti tomuto způsobu útoku není systém jakkoliv zabezpečen. Poslední část této kapitoly jsem okrajově věnoval softwaru USM Anywhere, který nabízí způsob, jakým dokáže administrátor sledovat kybernetické hrozby a zranitelnosti v systému.

SEZNAM POUŽITÉ LITERATURY

[1] *Wikipedia: Computer cluster* [online]. 2020 [cit. 2022-02-18]. Dostupné z: https://en.wikipedia.org/w/index.php?title=Computer_cluster&oldid=510200868

[2] Form 10-K: Annual report which provides a comprehensive overview of the company for the past year. *SEC Filing | Dell Technologies* [online]. United States, 31.1.2020 [cit. 2022-02-18]. Dostupné z: https://investors.delltechnologies.com/node/10741/html

[3] NVIDIA Corporation. Tensor Cores are the advanced NVIDIA technology that enables mixed-precision computing. This technology expands the full range of workload across AI & HPC. *NVIDIA Tensor Cores: Versatility for HPC & AI* [online]. United States: NVIDIA Corporation, 2022, 21.04.2022 [cit. 2022-02-18]. Dostupné z: https://www.nvidia.com/en-us/data-center/tensor-cores/

[4] HAN, Frank a Dharmesh PATEL. *Deep Learning Training Performance on Dell EMC PowerEdge R7525 Servers with NVIDIA A100 GPUs* [online]. 2020-11-11 [cit. 2022-02-18]. Dostupné z: https://infohub.delltechnologies.com/p/deep-learning-training-performance-on-dell-emc-poweredge-r7525-servers-with-nvidia-a100-gpus/

[5] Dell. *Rackový server Dell EMC PowerEdge R7525 | Dell Česká republika* [online]. 2019 [cit. 2022-02-18]. Dostupné z: https://www.dell.com/cz/domacnosti/p/poweredge-r7525/pd

[6] FUJITSU. *Company milestones : Fujitsu Global* [online]. [cit. 2022-02-20]. Dostupné z: https://www.fujitsu.com/global/about/corporate/history/company-milestones/

[7] FUJITSU. *FUJITSU Server PRIMERGY GX2570 M6 : Fujitsu Global* [online]. 2022 [cit. 2022-02-20]. Dostupné z: https://www.fujitsu.com/global/products/computing/servers/primergy/gpu/gx2570m6/#documents

[8] NVIDIA CORPORATION. *NVIDIA A100 TENSOR CORE GPU: Unprecedented Acceleration at Every Scale* [online]. June 2021, 3 [cit. 2022-02-21]. Dostupné z: https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf

[9] MAKU MAKU S.R.O. *Mining case 4U* [online]. Praha, 2021 [cit. 2022-02-23]. Dostupné z: https://www.miningcase.cz/p/mining-case#

[10] ASROCK INC. *ASRock X399 Taichi* [online]. 2022 [cit. 2022-02-23]. Dostupné z: https://www.asrock.com/mb/amd/x399%20taichi/index.asp#Specification

[11] LAMPORT, L. How to Make a Multiprocessor Computer That Correctly Executes Multiprocess Programs. *IEEE Transactions on Computers, Computers, IEEE Transactions on, IEEE Trans. Comput* [online]. 1979, **C-28**(9), 690-691 [cit. 2022-02-25]. ISSN 00189340. Dostupné z: doi:10.1109/TC.1979.1675439

[12] ADVANCED MICRO DEVICES, INC. *AMD Ryzen™ Threadripper™ 1920X* [online]. 2022 [cit. 2022-03-01]. Dostupné z: https://www.amd.com/en/product/2061

[13] ALCORN, Paul a Igor WALLOSSEK. *AMD Ryzen Threadripper 1920X Review* [online]. United States: Future US, 2017-08-30 [cit. 2022-03-07]. Dostupné z: https://www.tomshardware.com/reviews/amd-ryzen-threadripper-1920x-cpu,5183-10.html

[14] ZIVANOVIC, Darko, Milan PAVLOVIC, Milan RADULOVIC, et al. Main Memory in HPC. ACM Transactions on Architecture and Code Optimization [online]. 2017, 14(1), 1-26 [cit. 2022-03-12]. ISSN 1544-3566. Dostupné z: doi:10.1145/3023362

[15] QIN, Feng, Shan LU a Yuanyuan ZHOU. SafeMem: exploiting ECC-memory for detecting memory leaks and memory corruption during production runs. *11th International Symposium on High-Performance Computer Architecture, High-Performance Computer Architecture, 2005. HPCA-11. 11th International Symposium on, High-Performance Computer Architecture* [online]. 2005, 291-302 [cit. 2022-03-28]. ISBN 0769522750. ISSN 15300897. Dostupné z: doi:10.1109/HPCA.2005.29

[16] MICROWAY. *Introduction to RAID for HPC Customers* [online]. 2015-04-06 [cit. 2022-03-28]. Dostupné z: https://www.microway.com/hpc-tech-tips/introduction-raid-hpc-customers/

[17] KATZ, Randy H. RAID: A Personal Recollection of How Storage Became a System. *IEEE Annals of the History of Computing, Annals of the History of Computing, IEEE, IEEE Annals Hist. Comput* [online]. 2010, **32**(4), 82-87 [cit. 2022-04-13]. ISSN 10586180. Dostupné z: doi:10.1109/MAHC.2010.66

[18] STEWART, Samuel. *What Are NVIDIA CUDA Cores And What Do They Mean For Gaming?* [online]. 2022-01-10 [cit. 2022-04-13]. Dostupné z: https://www.gamingscan.com/what-are-nvidia-cuda-cores/

[19] NVIDIA CORPORATION. *Grafické karty NVIDIA GeForce RTX řady 30 poháněné architekturou Ampere* [online]. [cit. 2022-04-13]. Dostupné z: https://www.nvidia.com/cs-cz/geforce/graphics-cards/30-series/

[20] NVIDIA CORPORATION. *ŘADA GEFORCE RTX 3090 FAMILY* [online]. In: . [cit. 2022-04-15]. Dostupné z: https://www.nvidia.com/content/dam/en-zz/Solutions/geforce/ampere/rtx-3090/geforce-rtx-3090-shop-630-d@2x.png

[21] MCOMPUTERS. *NVIDIA ACCELERATORS FOR DATA CENTRES* [online]. [cit. 2022-04-16]. Dostupné z: https://mcomputers.cz/en/nvidia/accelerators/

[22] TECHNOSTORE LLC. *2021 2020 Deep Learning Benchmarks Comparison: NVIDIA RTX 3090 vs NVIDIA RTX A4000 | BIZON Custom Workstation Computers. Best Workstation PCs and GPU servers for AI, deep learning, video editing, 3D rendering, CAD.* [online]. [cit. 2022-04-17]. Dostupné z: https://bizon-tech.com/gpu-benchmarks/NVIDIA-RTX-3090-vs-NVIDIA-RTX-A4000/579vs604

[23] HE, Kaiming, et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 770-778.

[24] HENNESSY, John L., Krste ASANOVIĆ a David A. PATTERSON. *Computer Architecture: A Quantitative Approach*. 2012. ISBN 9780123838728. Dostupné z: https://search.ebscohost.com/login.aspx?direct=true&db=edsebk&an=407995&scope=site

[25] TECHNOSTORE LLC. *2021 2020 Deep Learning Benchmarks Comparison: NVIDIA RTX 3090 vs NVIDIA RTX A6000 vs NVIDIA RTX A4000 vs NVIDIA RTX A5000 | BIZON Custom Workstation Computers. Best Workstation PCs and GPU servers for AI, deep learning, video editing, 3D rendering, CAD.* [online]. [cit. 2022-04-19]. Dostupné z: https://bizon-tech.com/gpu-benchmarks/NVIDIA-RTX-3090-vs-NVIDIA-RTX-A6000-vs-NVIDIA-RTX-A4000-vs-NVIDIA-RTX-A5000/579vs585vs604vs605

[26] FUJITSU LIMITED. *Remote Management Controller User's Guide* [online]. [cit. 2022-04-20]. Dostupné z: https://www.fujitsu.com/global/Images/b7fh-5631-01en.pdf

[27] DELL. *Dell PowerEdge: How to configure the iDRAC & System Management Options on servers* [online]. 2021 [cit. 2022-04-21]. Dostupné z: https://www.dell.com/support/kbdoc/en-us/000179517/dell-poweredge-how-to-configure-the-idrac-system-management-options-on-servers

[28] LAYTON, Jeff. *Monitoring HPC Systems: What Should You Monitor?* [online]. [cit. 2022-04-21]. Dostupné z: https://www.admin-magazine.com/HPC/Articles/HPC-Monitoring-What-Should-You-Monitor

[29] ZABBIX LLC. *O Zabbix LLC* [online]. [cit. 2022-04-21]. Dostupné z: https://www.zabbix.com/cz/about

[30] ZABBIX LLC. *Zabbix Manual* [online]. [cit. 2022-04-21]. Dostupné z: https://www.zabbix.com/documentation/5.4/en/manual/

[31] ZABBIX LLC. *What's New in Zabbix 5.0* [online]. [cit. 2022-04-21]. Dostupné z: https://www.zabbix.com/whats_new_5_0

[32] *Wikipedie: Otevřená encyklopedie: IDRAC* [online]. c2021 [citováno 21. 04. 2022]. Dostupné z: https://cs.wikipedia.org/w/index.php?title=IDRAC&oldid=20057553

[33] DELL INC. *Integrated Dell Remote Access Controller 9 User's Guide* [online]. 2020 [cit. 2022-04-21]. Dostupné z: https://dl.dell.com/topicspdf/44010ug_en-us.pdf?fbclid=IwAR3tZ7Nw72JBsISMD92-lrLSbWrOEeNQnFWAep4wRP-gjyk6VcuMB5_15Flg

[34] KENNEDY, Patrick. *Dell iDRAC 8 Enterprise Overview: Excellent server management* [online]. 2016 [cit. 2022-04-22]. Dostupné z: https://www.servethehome.com/dell-idrac-8-enterprise-overview/

[35] ZABBIX LLC. Proxies. *Zabbix documentation* [online]. [cit. 2022-04-22]. Dostupné z: https://www.zabbix.com/documentation/current/en/manual/distributed_monitoring/proxies

[36] LAMBERT, Dmitry. *Zabbix SNMP – What You Need to Know and How to Configure It* [online]. 2020-06-18 [cit. 2022-04-22]. Dostupné z: https://blog.zabbix.com/zabbix-snmp-what-you-need-to-know-and-how-to-configure-it/10345/

[37] CHOLLET, François. *Deep learning v jazyku Python: knihovny Keras, Tensorflow*. Praha: Grada Publishing, 2019, 328 s. Knihovna programátora. ISBN 9788024731001.

[38] KERAS TEAM. *About Keras* [online]. [cit. 2022-04-22]. Dostupné z: https://keras.io/about/#installation-amp-compatibility

[39] MD. REZAUL KARIM. *Deep Learning with TensorFlow: Explore Neural Networks and Build Intelligent Systems with Python, 2nd Edition*. 2018. ISBN 9781788831109. Dostupné také z: https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&an=1789473&scope=site

[40] GOOGLE BRAIN TEAM. *Install TensorFlow 2* [online]. 2022-02-15 [cit. 2022-05-17]. Dostupné z: https://www.tensorflow.org/install

[41] NVIDIA CORPORATION. *CUDA GPUs* [online]. [cit. 2022-05-17]. Dostupné z: https://developer.nvidia.com/cuda-gpus

[42] ABADI, Martín, Paul BARHAM, Jianmin CHEN, et al. *TensorFlow: a system for large-scale machine learning* [online]. USA: USENIX Association, 2016 [cit. 2022-05-17]. ISBN 978-1-931971-33-1. Dostupné z: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

[43] VISHNU SUBRAMANIAN. *Deep Learning with PyTorch: A Practical Approach to Building Neural Network Models Using PyTorch*. 2018. ISBN 9781788624336. Dostupné také z: https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&an=1728031&scope=site

[44] CHOI, Jake, Heon Young YEOM a Yoonhee KIM. *2021 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C), Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C), 2021 IEEE International Conference on, ACSOS-C*. 2021, 20-25. ISBN 9781665443937. Dostupné z: doi:10.1109/ACSOS-C52956.2021.00029

[45] PYTORCH. *GET STARTED* [online]. [cit. 2022-05-18]. Dostupné z: https://pytorch.org/get-started/locally/

[46] LAMBDA. *Deep Learning GPU Benchmarks* [online]. [cit. 2022-05-18]. Dostupné z: https://lambdalabs.com/gpu-benchmarks

[47] CANONICAL LTD. *Ubuntu 20.04.4 LTS (Focal Fossa)* [online]. [cit. 2022-05-18]. Dostupné z: https://releases.ubuntu.com/20.04/

[48] NVIDIA CORPORATION. *CuDNN Archive* [online]. [cit. 2022-05-19]. Dostupné z: https://developer.nvidia.com/rdp/cudnn-archive

[49] PYTHON SOFTWARE FOUNDATION. *Ai-benchmark 0.1.2: AI Benchmark is an open source python library for evaluating AI performance of various hardware platforms, including CPUs, GPUs and TPUs.* [online]. 2019-12-08 [cit. 2022-05-19]. Dostupné z: https://pypi.org/project/ai-benchmark/

[50] A.I. *AI Benchmark for Windows, Linux and macOS: Let the AI Games Begin...* [online]. [cit. 2022-05-20]. Dostupné z: https://ai-benchmark.com/alpha.html

[51] A.I. *Deep Learning Hardware Ranking* [online]. [cit. 2022-05-20]. Dostupné z: https://ai-benchmark.com/ranking_deeplearning.html

[52] ANAFRA S.R.O. *NVIDIA Tesla V100 SXM2 32GB CoWoS HBM2, NVLink, GPU-NVTV100-32-SXM2* [online]. [cit. 2022-05-20]. Dostupné z: https://smicro.cz/nvidia-tesla-v100-sxm2-32gb-cowos-hbm2-nvlink-gpu-nvtv100-32-sxm2-1

[53] DR. KINGHORN, Donald. *Quad RTX3090 GPU Wattage Limited "MaxQ" TensorFlow Performance* [online]. 2020-11-13 [cit. 2022-05-20]. Dostupné z: https://www.pugetsystems.com/labs/hpc/Quad-RTX3090-GPU-Wattage-Limited-MaxQ-TensorFlow-Performance-1974/

[54] GOOGLE BRAIN TEAM. *Optimize TensorFlow performance using the Profiler* [online]. [cit. 2022-05-20]. Dostupné z: https://www.tensorflow.org/guide/profiler

[55] GOOGLE BRAIN TEAM. *Optimize TensorFlow GPU performance with the TensorFlow Profiler* [online]. [cit. 2022-05-21]. Dostupné z: https://www.tensorflow.org/guide/gpu_performance_analysis

[56] TEACHES TECH, Tony. *How to Enable Automatic Updates and Security Updates in Ubuntu* [online]. 2021-08-04 [cit. 2022-05-21]. Dostupné z: https://tonyteaches.tech/ubuntu-automatic-update-tutorial/

[57] TEACHES TECH, Tony. *10 Simple Ways to Secure Ubuntu from Hackers* [online]. 2021-05-18 [cit. 2022-05-21]. Dostupné z: https://tonyteaches.tech/secure-ubuntu-server/

[58] BOUCHERON, Brien a Justin ELLINGWOOD. *How To Configure SSH Key-Based Authentication on a Linux Server* [online]. 2021-06-16 [cit. 2022-05-21].

Dostupné z: https://www.digitalocean.com/community/tutorials/how-to-configure-ssh-key-based-authentication-on-a-linux-server

[59] HELPLOGICS. *Jak přidat klíče SSH na Ubuntu 20.04* [online]. 2022-12-02 [cit. 2022-05-22]. Dostupné z: https://helplogics.net/cs/jak-pridat-klice-ssh-na-ubuntu-20-04

[60] LINUXIZE. *How to Install and Configure Fail2ban on Ubuntu 20.04* [online]. 2020-08-19 [cit. 2022-05-22]. Dostupné z: https://linuxize.com/post/install-configure-fail2ban-on-ubuntu-20-04/

[61] AT&T CYBERSECURITY. *About OTX* [online]. [cit. 2022-05-22]. Dostupné z: https://cybersecurity.att.com/documentation/usm-anywhere/user-guide/otx/about-otx.htm?Highlight=OTX

[62] AT&T CYBERSECURITY. *USM Anywhere: Powerful threat detection and incident response for all your critical infrastructure* [online]. 2020 [cit. 2022-05-22]. Dostupné z: https://cdn-cybersecurity.att.com/docs/product-briefs/DS-USM-Anywhere.pdf

[63] AT&T CYBERSECURITY. *AlienVault threat intelligence* [online]. [cit. 2022-05-22]. Dostupné z: https://cybersecurity.att.com/solutions/threat-intelligence

[64] AT&T CYBERSECURITY. *USM Anywhere: Start detecting threats on day one and drive operational efficiency with one unified platform for threat detection, incident response, and compliance management.* [online]. [cit. 2022-05-22]. Dostupné z: https://cybersecurity.att.com/products/usm-anywhere

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

HPC	High performance computing
GPU	Graphics processing unit, grafický procesor
GB	Gigabyte
TB	Terabyte
SSD	Solid-state drive
Ghz	Gigahertz
nm	nanomentry
KB	Kilobyte
MB	Megabyte
CUDA	Compute Unified Device Architecture
W	Watt
Atd	A tak dále
SSL	Secure socket layer
LDAP	Lightweight Directory Access Protocol
IP	Internet Protocol
ks	kusy
AI	Artificial Intelligence
kWh	kilowatthodiny
SSH	Secure shell
IP	Internet protocol

SEZNAM OBRÁZKŮ

SEZNAM TABULEK