|  |  |
| --- | --- |
|  | |
| **Datově orientovaný přístup při vývoji software** | |
|  | |
| Bc. Tomáš Janečka | |
|  | |
|  |  |
| 2023 | Popis: fai_logo_cz |
|  |  |
|  | |

Text, letter

Description automatically generated

Text, letter

Description automatically generated**Prohlašuji, že**

* beru na vědomí, že odevzdáním diplomové 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 diplomová práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk diplomové práce bude uložen v příruční knihovně Fakulty aplikované informatiky Univerzity Tomáše Bati ve Zlíně;
* byl/a jsem seznámen/a s tím, že na moji diplomovou 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 – diplomovou 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 diplomové 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 diplomové práce využít ke komerčním  
  účelům;
* beru na vědomí, že pokud je výstupem diplomové 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 diplomové práci pracoval samostatně a použitou literaturu jsem citoval. V případě publikace výsledků budu uveden jako spoluautor.
  + že odevzdaná verze diplomové práce a verze elektronická nahraná do IS/STAG jsou totožné.

Ve Zlíně, dne TODO Tomáš Janečka, v.r.

ABSTRAKT

Diplomová práce se bude zabývat tématem datově orientovaného vývoje software. V teoretické části se student bude zabývat popisem jednotlivých technik návrhu software, charakteristikou používaného hardware a popisem vybraných aspektů ovlivňujících výkonnost implementovaných algoritmů. V praktické části budou srovnány jednotlivé způsoby implementace algoritmů včetně praktických příkladů doplněné o výstupy nástrojů pro výkonnostní testy a profilování.

Klíčová slova: datová orientace, C++, optimalizace, výkon, benchmarking, profilování, vyrovnávací paměť, operační paměť

ABSTRACT

This diploma thesis is concerned with the topic of data-oriented software design. In the theoretical part, the student is tasked with describing software design techniques, the characteristics of the hardware used and pointing out selected aspects that determine the performance of the implemented algorithms. The practical part shows the comparison of the different ways of algorithm implementations, along with examples including outputs from performance measuring tools and profilers.

Keywords: data-oriented, C++, optimization, performance, benchmarking, profiling, cache, memoryTODO

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

OBSAH

[Úvod 8](#_Toc120888191)

1. [TEORETICKÁ ČÁST 9](#_Toc120888192)

[1 Nadpis Error! Bookmark not defined.](#_Toc120888193)

[1.1 Podnadpis **Error! Bookmark not defined.**](#_Toc120888194)

[1.2 Podnadpis **Error! Bookmark not defined.**](#_Toc120888195)

[1.2.1 Podpodnadpis 13](#_Toc120888196)

[2 Nadpis Error! Bookmark not defined.](#_Toc120888197)

[2.1 Podnadpis **Error! Bookmark not defined.**](#_Toc120888198)

[2.1.1 Podpodnadpis **Error! Bookmark not defined.**](#_Toc120888199)

1. [Praktická část 27](#_Toc120888200)

[3 Nadpis Error! Bookmark not defined.](#_Toc120888201)

[3.1 Podnadpis **Error! Bookmark not defined.**](#_Toc120888202)

[3.2 Podnadpis **Error! Bookmark not defined.**](#_Toc120888203)

[4 Nadpis Error! Bookmark not defined.](#_Toc120888204)

[4.1 Podnadpis **Error! Bookmark not defined.**](#_Toc120888205)

[Závěr 28](#_Toc120888206)

[Seznam použité literatury 32](#_Toc120888207)

[Seznam použitých symbolů a zkratek 33](#_Toc120888208)

[Seznam obrázků 34](#_Toc120888209)

[Seznam tabulek 35](#_Toc120888210)

[Seznam Příloh 36](#_Toc120888211)

TODO – obsah poskládat manuálně

Úvod

Objektově orientované programování je jedno z nejrozšířenějších paradigmat mezi programátory. Mnoho z nich, včetně mě, se s ním setkali na úplném začátku své programátorské kariéry. Jedná se o velmi užitečný nástroj. Každý nástroj má ale svůj účel, a není určen k řešení každého problému. Jedním ze zásadních rozdílů mezi objektově orientovaným a datově orientovaným návrhem je ta věc, kterou tyto způsoby programování považují jako hlavní. Objektově orientovaný návrh se soustředí na vytvoření abstraktního, idealizovaného a také co nejobecnějšího modelu reálného problému. Naproti tomu datově orientovaný návrh považuje data za to nejvýznamnější. Přeci jenom, programy ve své podstatě slouží k transformaci dat z jedné podoby do druhé. [TODO acton2014] Zároveň se toto paradigma soustředí na charakteristiky hardware, na kterém náš software běží a dbá na efektivní využívání zdrojů. Z tohoto důvodu je tento způsob tvorby programů mimo jiné využíván v herním průmyslu. Právě herní vývojáři tvoří naučné podklady o tomto tématu, ve kterých často poukazují na podstatné nuance při souhře hardware a software. Tyto zdánlivé detaily však často mají zásadní vliv na rychlost běhu programu a také na využití operační paměti. Osobně jejich nápady považuji za velmi zajímavé, a proto by tato diplomová práce měla být shrnutím nejzásadnějších myšlenek tohoto paradigmata. V teoretické části si představíme definici DOP a popíšeme hlavní myšlenky. Rovněž je třeba se zabývat tématem mikroarchitektury počítače, jelikož jeho znalost je pro využití v této oblasti kritická. V praktické části bude představeno množství praktik, které lze aplikovat na tvorbu programu a budou porovnána běžná řešení často řešených problémů a také zhodnoceny výstupy výkonnostních testů a profilovacích nástrojů.

TODO

|  |  |
| --- | --- |
|  | TEORETICKÁ ČÁST |

# Datově orientované programování

V této sekci si přiblížíme pojem datově orientované programování a datově orientovaný návrh. Kromě dále uvedených definic je možné o tomto paradigmatu říci, že se jedná o způsob, jakým vyvíjet software. Zároveň ale může koexistovat s kódem, který byl napsán způsobem jiným ve stejném projektu. Tento nástroj se neváže ke konkrétní oblasti problémů či programovacích jazyků. V žádném případě se nejedná o něco, co by nebylo použito v minulosti, byť například pod jiným jménem. Ačkoliv nejde o nový koncept, samotný pojem „data-oriented“ se ve vývojářských kruzích začal vyskytovat teprve nedávno. I z tohoto důvodu je třeba při této diplomové práci využít omezené množství knih, které se zabývají tímto tématem, ovšem také větší množství záznamů přednášek z programátorských konferencí. Navíc si může člověk při studiu této problematiky všimnout, že každý řečník či autor si pod tímto pojmem představuje něco trochu jiného. Některé koncepty může zcela zanedbat a také může představit něco, o čem nikdo před ním nemluvil.

## Definice

„Datově orientovaný návrh je dovednost navrhnout software pomocí vývoje transformací pro data v řádné formě, kde řádná forma je řízena cílovým hardwarem a transformacemi, které na něm běží.“ [TODO dodmain]

„Datově orientovaný návrh si nechává napovědět daty, která jsou pozorovatelná nebo očekávaná. Na rozdíl od uvažování všech možných scénářů nebo plánování adaptability, využíváme nejpravděpodobnější vstupy pro nasměrování algoritmu. Na rozdíl od plánování rozšiřitelnosti je jednoduchý a má za cíl splnit svůj úkol.“ [TODO dodmain]

## Historický výskyt

Článek na téma „data-oriented“, který jako jeden z prvních použil tento termín a také měl za cíl seznámit čtenáře s touto myšlenkou, vyšel v roce 2009 v časopisu Game Developer. Jedná se o příspěvek od herního vývojáře o způsobu vývoje her v časopisu pro herní vývojáře. Není divu, že toto paradigma pramení právě z oblasti, kde je souhra software a hardware klíčem k úspěchu. V tomto článku autor pojednává o tom, jak by všudypřítomné objektově orientované programování mohlo být příčinou nízkého výkonu her. Je v něm uvedeno, na kterou věc se různé programovací přístupy soustředí a jak se od nich datově orientovaný přístup liší. Autor dále uvádí svůj výčet výhod tohoto přístupu, a to paralelizace, využití vyrovnávací paměti, modularita a jednoduchost testování. Závěrem jsou představeny rady, jak začlenit tento přístup do aktuálně vyvíjené aplikace a jak data získat a co je důležité sledovat. Autor se ještě vyjadřuje k tomu, že pro objektově orientovaný návrh rozhodně existuje místo a nechce ho démonizovat. Například „v systémech, které byly navrženy tímto způsobem nebo výkonově nekritické aplikace.“ [TODO llopis] Odkazy z tohoto příspěvku míří na Mika Actona a Jima Tilandera. Oba jsou vlastníky webů, kde v minulosti publikovali blogy na různá témata, které zjevně ovlivnily Noelův přístup k vývoji aplikací a her. Z toho je zřejmé, že myšlenky datově orientovaného návrhu pramení z prvních let druhého tisíciletí.

Jelikož se zde také klade důraz na vývoj software s ohledem na hardware, dalo by se říct, že byl tento přístup používán ještě mnohem dříve. V raných dobách výpočetní techniky byly, v porovnání s dnešní dobou, všechny dostupné prostředky velmi vzácné, a proto bylo nezbytné s nimi nakládat co nejefektivněji.

## Hlavní myšlenky

### Je to o datech

„Data jsou vše, co máme“. Všechny aplikace, co kdy byly napsány, slouží k poskytnutí výstupu v závislosti na vstupních datech. Grafické aplikace pracují s obrázky. Textové editory pracují s textem. Každá z nich očekává určitý formát dat. Ten může být velmi složitý, nebo velmi jednoduchý. Programátoři si také často neuvědomují, že instrukce jsou také data, protože se rovněž nachází v operační paměti. Za všech okolností je potřeba myslet na to, že data nikdy neexistují jen tak v éteru, ale pokaždé se nachází na nějakém hardware, ať už na virtuálním stroji, nebo konkrétním procesoru. [TODO dodmain]

### Data nejsou problémová doména

„Datově orientovaný návrh nezabudovává problém z reálného světa do kódu“. Na rozdíl od objektově orientovaného přístupu, datově orientovaný přístup obětovává čitelnost kódu, což nám umožňuje nezatěžovat počítač lidskými koncepty. Umístěním dat do třídy nám umožní dát těmto datům kontext, to ale může mít následek existence velkého množství dat, které spolu nesouvisí. Proto v tomto paradigmatu uvažujeme o datech jako o „faktech, o kterých můžeme uvažovat tak, jak potřebujeme pro získání výstupních dat v požadovaném formátu.“ [TODO dodmain]

### Statistika

„Data jsou typ, frekvence, množství, tvar a pravděpodobnost.“ [TODO dodmain]

Nejen vstupní data programu jsou zahrnuta do pojmu „data“. Data o datech mohou být stejně nebo i více významná. Mohou mít zásadní vliv na to, jak píšeme kód. Pokud máme k dispozici prvotní verzi funkční aplikace, která pracuje s produkčními daty, nebo disponujeme daty, která byla shromážděna libovolným způsobem, máme nyní více kontextu a dokážeme lépe uvažovat o problému, který řešíme. „Pokud porozumíme datům, porozumíme problému.“ Analýza dat může mít podobu prostého výpisu hodnoty proměnné. Stačí, když zvolíme libovolnou proměnnou, která nás zajímá a budeme sledovat vývoj jejích hodnot v čase. [TODO acton2014] Jelikož vývoj aplikací může být velmi náročný na čas a prostředky, určitě je rozumné investovat naše úsilí do 20 % kódu, je kterém je tráveno 80 % času a má zásadní vliv na výkon programu.

### Data se mění

„Datově orientovaný návrh může pozorovat změnu v architektuře aplikace porozuměním změnám v datech.“ Významná myšlenka je držet data a operace nad těmito daty zvlášť a neshlukovat je do tříd nebo jiných logických konstruktů. Díky tomu dokážeme lépe reagovat na změny a minimalizovat náklady potřebného přepisu kódu. [TODO dodmain]

## Využití

Jak již bylo zmíněno, datově orientovaný návrh se hojně využívá v herním průmyslu. Je to jedna z oblastí, kde se vývojáři snaží vytěžit co možná největší výkon ze své aplikace a zároveň musí respektovat omezení jednoho nebo více druhů hardware, na kterém bude běžet. Po seznámení s tímto paradigmatem a jeho hlavními myšlenkami a způsoby implementace je mi zřejmé, že použití tohoto přístupu má pro programátory jako jednotlivce i další zajímavé implikace. Jedna z nich je zařazení dalšího užitečného nástroje mezi své dovednosti. Jelikož pro aplikování datově orientovaného návrhu je důležitá znalost hardware, je programátor nucen se vzdělávat v oblasti počítačové architektury, mikroarchitektury, operačních systémů a strojového kódu. Zároveň je zde potenciál lepšího porozumění problému, který je zrovna řešen programátorem. Protože pokud porozumíme datům, porozumíme problému, může člověk zjistit řadu hodnotných informací a podle toho může v budoucnu vylepšit kód. Mezi specifické informace by se dala zařadit frekvence volání určitých funkcí, studium hodnot proměnných měnících se v čase nebo vypozorování vzorce opakování hodnot proměnných.

### Podpodnadpis



Obrázek . Ukázkový obrázek

Tabulka . Ukázková tabulka

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# další paradigmata

Programovací paradigma je způsob nebo styl, kterým píšeme kód. Nemusí se nutně vztahovat ke konkrétnímu programovacímu jazyku, i když různé programovací jazyky mají blíže k jednomu paradigmatu než ke druhému. V následující sekci se nachází srovnání paradigmata, kterým se zabývá tato diplomová práce, s dalšími paradigmaty, které se běžně používají, nebo začínají používat v aktuální době. Byla vybrána pouze ta, která jsou pro srovnání relevantní, protože mají například řešit nedostatky druhého.

## Objektově orientované programování

„Objektově orientované programování je o modelování systému jako kolekci objektů, kde každý objekt představuje určitý aspekt systému. Objekty obsahují jak funkce, tak data. Objekt poskytuje veřejné rozhraní, které je přístupné v kódu, a také obsahuje svůj privátní, vnitřní stav; ostatní části systému se nemusí zajímat o to, co se děje uvnitř objektu.“ [TODO mozillaOOP]

V rámci tohoto paradigmata se hojně využívá tříd. Třída je předpis pro vytváření instancí. Každá třída může mít vlastnosti, které charakterizují instance, a metody, které popisují chování. Mezi třídami mohou existovat vztahy. Jedním z nich je dědičnost, díky které jedna třída dědí vlastnosti a metody třídy druhé. [TODO mozillaOOP] Pokud rodičovská třída definuje metody jako virtuální, pak třídy, které od ní dědí, mohou měnit chování v závislosti na typu instance. Toto nám umožňuje mít kolekci objektů různého typu, ale pro interakci s nimi používat jednotný přístup. Ačkoliv se jedná o užitečnou věc, jsou to právě volání virtuálních funkcí, které v určitých případech mohou být příčinou zásadního zpomalení aplikace.

Zapouzdření je myšlenka vymezení veřejného rozhraní a zároveň schování detailů fungování objektu. V případě potřebné změny nám toto umožní změnit kód pouze na jednom místě, jelikož navenek je přístupné pouze veřejné rozhraní, které nebylo třeba měnit. [TODO mozillaOOP]

## Funkcionální programování

„Funkcionální programování je přístup k vývoji software, který používá ryzí funkce pro vytvoření udržitelného software. Jinými slovy se jedná o tvoření programů aplikací a kompozicí funkcí.“ [TODO wtfp]

Jak vyplývá z názvu, funkce je zde základní stavební jednotka. Kromě jejich běžného použití jsou funkce využívány taky jako proměnné, argumenty funkcí nebo návratové hodnoty funkcí. Na rozdíl od ostatních paradigmat se zde preferuje použití proměnných, jejichž hodnota se po deklaraci nemění. Základní myšlenky tohoto programovacího stylu pochází z matematického nástroje zvaného lambda kalkul, který byl popsán ve 30. letech minulého století Alonzo Churchem. Mezi funkcionální programovací jazyky se řadí Haskell, Erlang, Clojure, LISP, Scala a Elixir. Postupem času i běžné programovací jazyky zařazují do svého arzenálu nástroje, které pramení z tohoto způsobu programování. [TODO wtfp]

Ve funkcionálním programování je významné používání rekurze. V závislosti na implementaci toto může způsobit značné zpomalení aplikace v porovnání s implementací pomocí použití klasické iterace ve smyčce.

## Porovnání

„Objektově orientovaný návrh je soustředěn na problém a jeho řešení. Objekty, abstraktní reprezentace věcí, které tvoří návrh řešení problému představeného v návrhovém dokumentu aplikace. Objekty manipulují pouze s těmi daty, které jsou potřeba pro jejich reprezentaci bez jakéhokoliv ohledu na hardware nebo na data z reálného světa nebo jejich množství. Z tohoto důvodu nám objektově orientovaný návrh umožní rychle sestavit první verze aplikací a tím pádem také první podobu kódu. Datově orientovaný návrh se k problému staví jinak. Na rozdíl od předpokladu, že nevíme nic o hardware, usuzujeme, že nevíme nic o řešeném problému.“ [TODO dodmain]

Rozdíl mezi objektově orientovaným a funkcionálním přístupem je ten, že zatímco OOP využívá imperativní přístup, který spočívá ve specifikaci kroků potřebných k vyřešení problému, FP využívá deklarativní přístup, který pracuje s výsledkem operace, nehledě na to, jak jsme k němu přišli. Dalším rozdílem je využití proměnných a konstantních proměnných. V FP se v případě přepisu vytvoří zcela nová proměnná, do které se překopíruje původní hodnota. Díky tomu se kód v případě potřeby snáze mění a lépe testuje a lépe se v něm hledají chyby. Autor doporučuje využívat OOP pro standardizované a přímočaré projekty a FP pro aplikace, které je třeba škálovat a musí být flexibilní. [TODO [wtfp](https://www.codingdojo.com/blog/what-is-functional-programming)]

Jedna věc, kterou sdílí funkcionální programování s datově orientovaným programováním je skutečnost, že obě paradigmata identifikovala nedostatky objektově orientovaného programování. Společná překážka je potřeba uvažovat zcela odlišně, než jak člověk myslel doposud při psaní programu. Pro programátora odchovaného na objektově orientovaném návrhu mohou být myšlenky aplikované v těchto alternativních programovacích přístupech relativně složité, ba i zpočátku nepochopitelné. Obě paradigmata se mohou vyskytovat po boku objektově orientovaného kódu ve stejné aplikaci a být použity tam, kde dávají smysl. Zatímco funkcionální návrh si klade za cíl zlepšit robustnost a modularita aplikace, datově orientovaný přístup se soustředí jak na modularitu, tak na zvýšení výkonu programu.

# hardware jako platforma

Datově orientovaný vývoj dbá na efektivní využívání hardware. V této sekci si popíšeme části počítače a jednotlivých komponentů, které mají zásadní vliv na rychlost běhu programu.

TODO – CPU memory gap (Hennessy & Patterson, Computer Architecture: A Quantitative Approach, 4th ed.)

## Vyrovnávací paměť

Vyrovnávací paměť (cache) je rychlé paměťové zařízení s malou kapacitou, které slouží k ukládání malých částí dat z paměťových zařízení nižších úrovní paměťové hierarchie. Pokud se bavíme o cache pamětech, máme zpravidla namysli uložiště, které se nachází na procesorovém čipu a je k dispozici výpočetním jádrům. Jako vyrovnávací paměť ale můžeme považovat i operační paměť ve vztahu k hard-disku nebo SSD. [TODO csprogrammer]

Cache paměť pro ukládání dat využívá technologie SRAM. V moderních procesorech se vyskytuje v několika úrovních. Každá úroveň má různou velikost a přístupovou dobu. Nejblíže k výpočetnímu jádru je úroveň L1. Tu ještě výrobci CPU separují na paměť pro instrukce a data, nazývané i-cache a d-cache. Tato úroveň bývá privátní pro jedno jádro. Na další úrovni se nachází úroveň L2. Ta je unifikovaná, takže obsahuje jak instrukce, tak data. Může být privátní pro jedno jádro nebo sdílená mezi všemi jádry. Nejvýše postavená je úroveň L3. Ta je společná pro všechna jádra a má největší kapacitu. [TODO csprogrammer]

Vývoj vyrovnávacích pamětí je jeden z přispěvatelů k relativně obrovským výkonům dnešních CPU. Tento růst bohužel mnohonásobně převyšuje ten u operačních pamětí. I z tohoto důvodu by se měli programátoři naučit, jak efektivně využívat cache paměť. [TODO caqa]

Chart, line chart, scatter chart

Description automatically generated

Obrázek : Vývoj doby trvání přístupu do procesorové a operační paměti. [TODO caqa]

### Struktura a adresování

Cache paměti uspořádávají data do sad (set). Každá sada obsahuje jeden nebo více řádků (line). Jeden řádek obsahuje kromě samotných dat, kterým se také říká blok (block), obsahuje ještě jeden bit *v*,který signalizuje validitu záznamu, a také štítek (tag), které se používají při adresování a načítání dat z cache. [TODO csprogrammer]

K adresování do vyrovnávací paměti se využívá celá adresa dat rozdělená na části. Počet bitů každé části je dán počtem bitů, které počítač používá pro adresování *m* (typicky 32 nebo 64 bitů), počtem sad *S* a velikostí bloku *B* (typicky 64 B). Parametry *S*a *B* jsou stanoveny výrobcem a každá konfigurace má své výhody i nevýhody. Abychom mohli získat počet bitů použitých pro tag *t*, vypočítáme počet bitů pro index sady a počet bitů pro offset v bloku . Počet bitů pro tag . [TODO csprogrammer]

A picture containing background pattern

Description automatically generated

Obrázek : Příklad rozdělení adresy. Zdroj vlastní.

Bity *s* slouží pro nalezení odpovídající sady v cache paměti. Bity *t* slouží pro identifikaci odpovídající cache line v sadě. Bity *b* nám říkají, kde v rámci je bloku začátek dat, o která žádáme. [TODO csprogrammer]

### Pojmy

Při popisování funkce vyrovnávacích pamětí je třeba si popsat několik běžně používaných pojmů. *Cache hit* označuje nalezení požadovaných dat v první sousední úrovni směrem dolů v paměťové hierarchii. Při optimalizaci programů je snaha uchovat data, se kterými se operuje, v jakékoliv úrovni cache na procesoru, jelikož i přístup do L3 je rychlejší než přístup do operační paměti. *Cache miss* označuje absenci požadovaných dat v první sousední úrovni směrem dolů v paměťové hierarchii. Je nutné, aby byla data načtena z úložiště a uložena do cache. Pokud je cache paměť plná, je třeba využít zvolené substituční strategie a nahradit nějakou cache line. Toto může být označeno jako *block eviction*. Existuje několik druhů *cache miss*. V případě, že je vyrovnávací paměť prázdná, při žádosti o data nastává *compulsory miss*. Pokud by velikost vyrovnávací paměti byla příliš malá, nebo bychom se do ní pokoušeli zapisovat data, která musí být uložena na stejné místo, jedná se o *conflict miss*. Jestliže pracujeme obrovskou sadou dat, iterujeme přes ni ve smyčce a tím dojde k vyčerpání kapacity cache, pozorujeme *capacity miss*. Když dochází k opakovaným konfliktům na stejném místě v cache paměti, označujeme to jako *thrashing*. Poměr počtu *cache miss* a počtu dotazů na data nám dává *miss rate*. *Hit rate* je vyjádřeno jako . Doba přesunu dat z paměti do CPU se označuje jako *hit time*. Doba čekání na data v případě *cache miss* se jmenuje *miss penalty*. Bavíme-li se o pojmu *write hit*, máme na mysli to, že se data, která chceme aktualizovat, vyskytují v nejbližší cache paměti. Naproti tomu *write miss* označuje absenci dat, která aktualizujeme. [TODO csprogrammer]

### Vyrovnávací paměť s přímým mapováním

Direct-mapped cache je typ vyrovnávací paměti, který má v každé sadě právě jednu cache line. Tyto paměti jsou jednoduché na implementaci a používání, ale je zde zvýšené riziko *thrashingu*. [TODO csprogrammer]

### Set associative vyrovnávací paměť

Tento typ pamětí se redukuje problém existující v paměti s přímým mapováním pomocí navýčení počtu cache line v rámci jedné sady. *Asociativní* paměť ukládá data jako pole dvojic *klíč-hodnota*. Každou sadu v této paměti si lze představit jako asociativní paměť, jejímž klíčem je spojení bitů *v* a *t* a hodnotou je obsah bloku. [TODO csprogrammer]

Při návrhu stupně asociativity je třeba zvolit vhodný kompromis mezi větším a nižším. Vyšší stupeň asociativity je pomalejší a složitější na implementaci. [TODO csprogrammer]

Table

Description automatically generated

Obrázek : Příklad uspořádání a typu cache pamětí – snímek programu CPU-Z. Zdroj vlastní.

Zde je příklad různých úrovní cache pamětí. V levém sloupci vidíme *počet jader x velikost paměti*. V pravém sloupci je uveden stupeň asociativity. *N-way* nám říká, že každá sada pojme *N* cache line.

### Plně asociativní vyrovnávací paměť

Jedná se o asociativní vyrovnávací paměť s jednou sadou. Tato sada obsahuje všechny cache line. Při adresování je možné vypustit bity *s*, jelikož se vždy pracuje s hodnotou 0. Plně asociativní vyrovnávací paměť je vhodná pro malé kapacity, jelikož pro větší kapacity by bylo zapotřebí značné množství hardware a zároveň by byla pomalá. [TODO csprogrammer]

### Substituční strategie

Pokud nastane *conflict miss*, je třeba umístit žádaná data na vhodné místo. Toto místo může být zvoleno náhodně. Alternativně můžeme zvolit sofistikovanější postupu, jako je nahrazení dat, která byla použita nejdále v minulosti (LRU), případně nejméně často využita ve stanoveném časovém okně (LFU). [TODO csprogrammer]

### Způsoby zapisování

Na rozdíl od čtení dat, zápis dat je o něco složitější. Jsou-li data, která aktualizujeme, v cache paměti, můžeme buď nová data rovnou zapsat do cache paměti nižší úrovně (*write-through*) nebo upravit data a zapsat je do paměti až v době vyřazení cache line z paměti (*write-back*). Nastane-li *write miss*, máme opět na výběr. Způsob *write-allocate* si nejdříve načte blok z nižší paměťové vrstvy a poté upraví data. Na rozdíl od toho, *no-write-allocate* přímo zapíše data do nižší úrovně. Výše uvedené způsoby lze kombinovat mezi sebou a každá kombinace je vhodná pro jiný cíl. [TODO csprogrammer]

## Operační paměť

Operační paměť na rozdíl od vyrovnávacích pamětí využívá technologie DRAM pro ukládání dat. Jeden bit je uchován pomocí kondenzátoru. Kvůli pokročilému stupni integrace a zároveň náchylnosti na rušení, každý kondenzátor musí být pravidelně dobíjen (refresh). Tento způsob ukládání dat je využit pro operační paměti kvůli možnosti dosažení vyšší kapacity na stejnou plochu čipu v porovnání s SRAM a také kvůli ceně. Čas přístupu je ale delší. [TODO csprogrammer]

Data na paměťovém čipu jsou uspořádána do dvourozměrného pole označovaného jako superbuňka (supercell) a ta zase obsahuje určitý počet paměťových buněk (cell). Čip je propojen s paměťovým kontrolérem pomocí adresových a datových vodičů, které slouží k vyhledání požadovaných buněk a zápisu nebo čtení dat. Při adresování se nejprve na vodiče zapíše adresa řádku, který se následně celý zkopíruje do interního bufferu řádku, a následně se na stejné vodiče zapíše adresa sloupce, což zapříčiní zapsání dat z požadované paměťové lokace z bufferu řádku na datové vodiče. [TODO csprogrammer]

Paměťové čipy jsou dále uspořádány do paměťových modulů. Přístup jednotlivým čipům na modulu probíhá paralelně. V případě čtení dat z paměti paměťový kontrolér obdrží adresu, kterou rozdělí podle počtu čipů na modulu pro získání adresy superbuňky. Každý čip zapíše hledaná data zpět na výstup modulu, který data poskládá do správného pořadí a poté je pošle kontroléru. Data jsou pomocí sběrnice následně zaslána do CPU. [TODO csprogrammer]

## CPU pipelining

### Fáze instrukčního kanálu

Zpracování jedné instrukce procesorem je rozděleno do několika fází a bývá popsáno instrukční *pipeline*. První je fáze *fetch*, ve které se načte instrukce z paměti pomocí adresy, která je uchována v registru jménem *program counter*. Následující úroveň *decode* se zabývá načtením operandů. Krok *execute* provede operaci, kterou popisuje instrukce, pomocí aritmeticko-logické jednotky. Může se jednat o matematické nebo logické operace, výpočet adresy, vyhodnocení podmínky nebo směr větvení. Zápis a čtení paměti se děje ve fázi *memory*. Konečný krok *write-back* zapisuje vypočtené výsledky do registrů. Abychom mohli hodnotit vhodnost různých přístupů k provádění instrukcí, používáme pojmy *latency*, což je doba vykonání operace, a *throughput*, který popisuje počet provedených operací za jednotku času. [TODO csprogrammer]

### Sekvenční zpracování

Jestliže vykonáváme instrukce jednu po druhé a každá instrukce projde kanálem jako celek, hovoříme o sekvenčním vykonávání.

Diagram, table

Description automatically generated

Obrázek : Sekvenční vykonávání instrukcí. Zdroj vlastní.

### Skalární pipeline

Jiný způsob provádění instrukcí, který by měl lépe využívat dostupný hardware, se nazývá *pipelined*.

Table

Description automatically generated with medium confidence

Obrázek : Skalární instrukční pipeline. Zdroj vlastní.

### Superskalární pipeline

Vylepšení skalárního *pipelined* přístupu je způsob superskalární. Ten je zdokonalen pomocí přidaného hardware, díky kterému se může během jednoho taktu nacházet více instrukcí v té samé fázi.

Table

Description automatically generated

Obrázek : Superskalární instrukční pipeline. Zdroj vlastní.

### Ideální vs reálné řešení

Ačkoliv je *pipelining* vhodným způsobem pro zvýšení počtu vykonaných instrukcí za jednotku času, může mít za následek mírný nárůst latence každé instrukce. Ve dříve popisovaných příkladech *pipeline* jsou uvažovány uniformní časy strávené v každé fázi. Ve skutečnosti však mohou různé kroky trvat různě dlouho a tím pádem způsobovat prodlevy při přechodu mezi kroky. Toto může být částečně vyřešeno zvýšením počtu fází a snížení doby trvání každé fáze, což, jako každé rozhodnutí při návrhu hardware, má taky své nevýhody. [TODO csprogrammer]

### Datové a kontrolní závislosti

Doposud byla uvažována nezávislost dvou po sobě jdoucích instrukcí. V reálném světě k takovému ideálnímu případu nemusí docházet často. Rozlišujeme *datovou závislost*, kde se dvě instrukce odkazují na to samé místo v paměti nebo ten samý registr, a *kontrolní závislost*, ke které dochází při skoku nebo návratu z rutiny. Nesprávná implementace řešení těchto stavů může zapříčinit chyby ve výpočet, které bývají označovány jako *data* a *control hazard*. [TODO csprogrammer]

Table

Description automatically generated

Obrázek : Data hazard. Zdroj vlastní.

*Data hazard* může mít podobu načtení staré hodnoty aktuální instrukcí, zatímco předchozí instrukce už měla hodnotu změnit na novou. Obrázek 7 ukazuje vznik chyby ve fázi D u instrukce I4. Jelikož je hodnota R1 aktualizována ve stejném taktu, jako je načítání hodnot operandů instrukce, načtená hodnota není správná. Tento problém lze vyřešit vícero způsoby. Ten nejjednodušší je pozastavení (stalling) pipeline. Spočívá ve vložení takzvaných bublin (bubble) mezi problémové instrukce, aby se posunuly překrývající fáze v rámci cyklu a tím pádem byla načtena správná hodnota. Tento přístup může mít negativní vliv na počet provedených instrukcí za jednotku času. Další způsob se nazývá *data forwarding*. Namísto toho, abychom nová data zapisovali ve fázi W do souboru registrů, aby mohla být v dalším taktu načtena fází D, můžeme aktualizovanou hodnotu předat přímo z W na požadované místo. [TODO csprogrammer]

Pokud ve fázi F nedokážeme jednoznačně určit adresu další instrukce, mluvíme o *control hazard*. Příkladně u instrukce RET, po jejímž zavolání by se obecně měl vrátit proud vykonávání na místo předchozího volání funkce, je adresa další instrukce známá až ve fázi W a efektivně tedy dochází k pozastavení pipeline. V případě instrukce podmíněného skoku se nemusí čekat na výsledek porovnání, ale lze odhadnout, kterým směrem větve se bude pokračovat. Jen tehdy, pokud se nesprávně zvolená instrukce nachází ve fázích F nebo D je můžeme po detekci bez problémů odstranit z pipeline, protože ještě neměly čas změnit aktuální kontext. [TODO csprogrammer]

## Prefetching

Prefetching je jeden ze způsobů tolerance latence načítání dat z paměti. Zabývá se načítáním dat z paměti předtím, než jsou potřeba, například pomocí predikce adres. Tímto způsobem můžeme eliminovat *compulsory cache miss*. V moderních systémech se velikost dat, která jsou přednačtena, rovná velikosti cache bloku. Tato technika může být provedena hardwarově, na úrovni kompilátoru nebo uživatelsky. Existuje více druhů prefetcherů a ty se mohou lišit tím, který přednačítací algoritmus je použit. Přednačítací algoritmus udává, co se bude načítat. Většina moderních systémů ukládá přednačtená data do cache paměti. [TODO mutluPrefetch]

### Softwarový prefetching

Prefetching pomocí software je proveden instrukcemi pro přednačítání dat, které poskytuje instrukční architektura. Pracuje s nimi programátor nebo kompilátor a jsou vhodné pro přístupy do paměti, které jsou pravidelné. Je možné vybrat si, do které úrovně cache paměti budou data zapsána. [TODO mutluPrefetch]

### Hardwarový prefetching

Prefetching pomocí hardware spočívá v monitorování přístupů do paměti a nalezení opakujících se vzorů. Předpovězené adresy jsou generovány automaticky. Na rozdíl od softwarového přístupu, hardwarový způsob může být lépe přizpůsoben systému, který používáme, a také může představovat menší zatížení. Funkce takového prefetcheru může být taková, že při čtení určitého cache bloku z paměti automaticky načte i jeden nebo více dalších, které po něm následují (*next-line prefetcher*). Jestliže je zaznamenáno několik opakujících se instrukcí, které načítají z paměti s pravidelným offsetem od počáteční adresy, může v určitém případě dojít k přednačtení dalších pravděpodobných dat (*stride prefetcher*). [TODO mutluPrefetch]

### Execution based prefetching

Univerzální způsob, který může být implementován jak hardwarově, tak softwarově, je *execution based prefetching*. Zde probíhá přednačítání dat v jiném vlákně, než ve kterém je prováděn hlavní program. [TODO mutluPrefetch]

Ve speciálním vlákně stačí provést ty části kódu, při kterých by jinak nastal *cache miss*. Mezi ně patří výpočet adres, předpovídání větvení a předpovídání hodnoty. Toto vlákno může existovat jak na stejném jádře, jako je vykonávaný kontext, tak na zcela separátním jádře. Rovněž nemusí existovat po celou dobu běhu programu, ale může dynamicky vznikat a zanikat. Pokud to ISA nabízí, může programátor využít poskytované instrukce pro zahájení výpočtu problémového bloku kódu. [TODO mutluPrefetch2]

### Address correlation based prefetching

Tvrzení, že se snáze odhadují pravidelně se posouvající přístupy do paměti, není překvapující. Ne u všech datových struktur, použitých pro práci s daty, je toto možné. Proto existují i různé přístupy k prefetchingu, které se snaží zrychlit i nepravidelné přístupy do paměti. Tento způsob využívá korelační tabulku pro zaznamenání přímého přístupu do paměti, a zároveň také těch, které následovaly hned po něm. Lze totiž vypozorovat, že některé přístupy mohou nastat s větší pravděpodobností než ostatní. Toto lze znázornit Markovovým modelem. Ke každému odkazu do paměti může být zapamatován jeden nebo více dalších následujících odkazů a ty budou v budoucnu přednačteny. Tento druh prefetchingu ovšem nedokáže eliminovat *compulsory cache miss*. [TODO mutluPrefetch2]

### Content directed prefetching

Při programování v jazycích jako C/C++ se často setkáváme s prací s ukazateli do paměti, zejména pak u některých datových struktur. Jelikož ukazatel se může odkazovat na jakékoliv místo v paměti, jen těžko bychom mohli využít předpovídání adres s pravidelným krokem. Právě pro zefektivnění práce s ukazateli řeší tento přístup. Snaží se identifikovat hodnoty v cache bloku, které jsou potencionálními ukazateli, a načíst pro ně odkazovanou hodnotu z paměti do cache. Není zde třeba žádné tabulky pro sledování přístupů v minulosti, ale také je zde možnost zahlcování cache paměti nepotřebnými daty, protože tento algoritmus načte data pro všechny ukazatele v cache bloku. [TODO mutluPrefetch2]

## Lokalita

Pojem lokalita se týká odkazování na data v paměti. Je to vlastnost dobře napsaných programů, které díky práci s ní mohou běžet rychleji, a to díky faktu, že hardware počítače i operační systémy jsou navrženy pro zužitkování lokality. Můžeme se s ní setkat v různé podobě v každé úrovni paměťové hierarchie. Zásadní myšlenka, která mohla podnítit vznik vyrovnávacích pamětí, byla, že pokud se jednou odkážeme na určité místo v paměti, je zde pravděpodobnost, že se na něj odkážeme vícekrát v blízké budoucnosti. Toto bývá označování jako *časová lokalita*. Jiným případem je *prostorová lokalita*, která říká, že pokud se odkážeme na určité místo v paměti, nejspíše se budeme odkazovat i na další místa v paměti v blízkosti původního. [TODO csprogrammer]

## Jazyk symbolických adres

V dnešní době masivního rozšíření high-level programovacích jazyků, jako je C#, Java, Python či JavaScript, se snáze zapomene na to, jak hardware počítače vlastně vykonává náš program. Je třeba si uvědomit, že pohodlnost používání těchto jazyků je výhodná zejména pro jeho uživatele. CPU počítače však netíží abstrakce, které do návrhu vnáší programátor, čímž i do určité míry zakrývá realitu toho, co se skutečně pod pokličkou děje. Skutečnost je taková, že procesor rozumí pouze konečné sadě instrukcí a umí je vykonat rychle. Nejblíže této úrovni, nebereme-li v potaz strojový kód, je jazyk symbolických adres. Porozumět kódu, který jsme napsali, v podobě jazyku symbolických adres je velmi užitečné a v kombinaci s pochopením architektury počítačového systému se jedná o velmi mocný nástroj při optimalizaci kódu.

Počítače vykonávají strojový kód, což je sekvence bytů popisující operace pro manipulaci s daty, čtení a zápis dat nebo síťovou komunikaci. Z high-level jazyku vznikne strojový kód pro konkrétní stroj s použitím kompilátoru. V jazyce symbolických adres programujeme pomocí skládání instrukcí za sebe jednu po druhé. Dostupné instrukce jsou definovány abstrakcí nazývanou *instruction set architecture*. Jedny z těch nejvyužívanějších jsou x86-32 a x86-64. [TODO csprogrammer]

### Vnitřní architektura x86-32

## Kompilátory

## SIMD

## Paralelizace

# měření výkonu aplikací

## Benchmarking

## Profilování

### Druhy profilerů

|  |  |
| --- | --- |
|  | Praktická část |

# příklady optimalizací

## Iterace polem

## Předvídatelnost operací

## Hot vs cold data

## SoA vs AoS

## Využití pipeliningu

## Zarovnání dat

## Datové a kontrolní závislosti

# ověření výkonu pomocí nástrojů

# souhrn optimalizací

Závěr

text

Seznam použité literatury

1. Text

TODO – očíslovat a smazat

Acton2014 – CppCon 2014: Mike Acton "Data-Oriented Design and C++". *Youtube* [online]. 30.9.2014 [cit. 2022-12-21]. Dostupné z: <https://www.youtube.com/watch?v=rX0ItVEVjHc>. Kanál uživatele CppCon

Dodmain – LLOPIS, Noel. Data-Oriented Design (Or Why You Might Be Shooting Yourself in The Foot With OOP) [online]. 2009-12-04 [cit. 2022-12-23]. Dostupné z: <https://gamesfromwithin.com/data-oriented-design>

mozillaOOP - Object-oriented programming [online]. 2022-09-28 [cit. 2022-12-30]. Dostupné z: <https://developer.mozilla.org/en-US/docs/Learn/JavaScript/Objects/Object-oriented_programming>

wtfp - MITCHELL, Brad. What Is Functional Programming and Why Use It? [online]. 2022-07-13 [cit. 2022-12-30]. Dostupné z: <https://www.codingdojo.com/blog/what-is-functional-programming>

csprogrammer – BRYANT, Randal a David O'HALLARON. Computer Systems: A Programmer's Perspective. 3rd Edition. Pearson, 2015, 1128 s. ISBN 013409266X.

mutluPrefetch – TODO <https://www.youtube.com/watch?v=ibPL7T9iEwY>

mutluPrefetch2 – TODO <https://www.youtube.com/watch?v=TUFins4z6o4>

caqa – HENNESSY, John L. a David A. PATTERSON. Computer Architecture: A Quantitative Approach. 4th Edition. 2006, 704 s. ISBN 0123704901.

asm - KUSSWURM, Daniel. Modern X86 Assembly Language Programming: 32-bit, 64-bit, SSE, and AVX. Apress, 2014, 700 s. ISBN 1484200659.

TODO – zkontrolovat na <https://odevzdej.cz/>

Seznam použitých symbolů a zkratek

|  |  |  |
| --- | --- | --- |
| DOD |  | Data-oriented design |
| DOP |  | Datově orientované programování |
| FP |  | Funkcionální programování |
| OOP |  | Objektově orientované programování |
| SSD |  | Solid State Drive |
| SRAM |  | Static Random Access Memory |
| CPU |  | Central Processing Unit |
| LRU |  | Least Recently Used |
| LFU |  | Least Frequently Used |
| DRAM |  | Dynamic Random Access Memory |
| ISA |  | Instruction Set Architecture |

Seznam obrázků

[Obrázek 1. Ukázkový obrázek 13](#_Toc56699409)

Obrázek 2: Příklad rozdělení adresy…………………………………………………13

TODO

Seznam tabulek

[Tabulka 1. Ukázková tabulka 13](#_Toc56699398)

Seznam Příloh

Příloha P I: Název přílohy