Faculdade projeção

DOD para otimização server-side em jogos

Documento de Visão

Versão 1.4.0

Histórico de Revisão

Data	Versão	Descrição	Autor
06/09/2020	1.0.0	Criação do Documento de Visão para descrição geral através do tema “DOD para otimização server-side em jogos”.	Nádio Dib
09/09/2020	1.1.0	Adição de conteúdo no capítulo 1 e revisão da formatação dos elementos textuais e contextuais deste item mediante adaptação para pesquisa científica.	Nádio Dib
09/09/2020	1.1.1	Adição de itens, definições e referências bibliográficas ao glossário deste documento. Além de correção de espaçamento de informações fornecidas no conteúdo do capítulo 1.	Nádio Dib
12/09/2020	1.2.0	Adição de conteúdo no capítulo 2.	Nádio Dib
12/09/2020	1.3.0	Adição de conteúdo no capítulo 3 e revisão da formatação dos elementos textuais do capítulo 2. Além de adição de itens pendentes no glossário.	Nádio Dib
12/09/2020	1.3.1	Correção ortográfica no título do capítulo 4.	Nádio Dib
12/09/2020	1.4.0	Adição de conteúdo no capítulo 4. Revisão da formatação dos elementos textuais dos capítulos 2 e 3. Correção de formatação de elementos das referências bibliográficas.	Nádio Dib

Sumário

Introdução

Ao decorrer de vários anos de trabalho informal na área de desenvolvimento de jogos para criação, manutenção e otimização de aplicações server-side ou *back-end* para servidores através da utilização do *framework* .NET, percebi que sobrecargas de execução nas operações de rotina repetitiva ou *loop* eram de intensa interação entre instâncias de objetos quando se tratava de comunicação por meio de encapsulamento de mensagens com a abordagem do paradigma *Object-Oriented Programming* (OOP) eram muito constantes e necessitavam e atenção.

Logo, a proposta deste documento é coletar, analisar e definir as necessidades e funcionalidades gerais do projeto "DOD para otimização server-side em jogos" para fins acadêmicos e de estudo prático.

Seu foco está nas necessidades da devida compreensão de conceitos e fundamentos computacionais para uma melhor adequação de operações exaustivas exigidas por aplicações de servidor de jogos através de uma nova abordagem feita pelos materiais levantados de DOD e os motivos da existência destas necessidades para mitigação dos problemas contextuais oriundos deste mesmo tema.

Termos e abreviaturas específicos podem ser encontrados no Glossário do respectivo projeto.

Contextualização

Dentro desta perspectiva, rotinas repetitivas em servidores de jogos tornaram-se regiões críticas e sensíveis durante suas operações quando em funcionamento no modo de produção ou *release*. Estas rotinas se caracterizaram como regiões críticas devido a implementação de conceitos de *multithreading* para utilização de recursos de programação paralela concorrente, nos quais, muitas das vezes eram caracterizados por recursos compartilhados entre processos adjuntos das rotinas para fins de otimização.

A utilização de conceitos e fundamentos dos assuntos de *concurrent programming* e *parallel programming* auxiliaram até de certa forma na confecção de uma solução para mitigação de sobrecargas de trabalho nestas operações exigidas por essas rotinas, que se caracterizavam por requisitar alto uso de recursos da CPU, no qual a aplicação estava em funcionamento, naquele ambiente computacional.

Entretanto, após investigar com mais detalhe sobre como possibilitar uma melhor adequação para estas rotinas, a simples utilização dos fundamentos de *Data-Oriented Design* (DOD) juntamente com o devido discernimento do funcionamento do sistema operacional, possibilitavam menor utilização dos recursos computacionais, pelo qual a aplicação está em funcionamento naquele ambiente operacional, e proporcionar confidencialidade no tratamento de dados para procedimentos de concorrência dos recursos, pois estes problemas recorrentes da utilização de conceitos e implementações de funcionalidades *multithreading* seriam ausentes, nos quais procedimentos assíncronos passariam a ser síncronos e pertencentes da mesma *thread* que caracteriza rotina.

De acordo com os assuntos abordados previamente na Introdução deste Documento de Visão e Escopo (DVE) de forma superficial, há a necessidade de apresentar esta abordagem oriunda de conceitos existentes das áreas de desenvolvimento de software como Sistemas Operacionais, Arquitetura de Software, Estrutura de Dados e Algoritmos, Modelagem de Software Orientado a Objeto e Lógica de Programação, nos quais fazem parte da formação acadêmica do curso de Engenharia de Software.

Problemática

Operações em rotinas que desempenham processos de longa duração e iterações entre inúmeras instâncias de objetos através do paradigma OOP, não são eficientes quando suscetível a uma grande coleção de instâncias de objetos para iteração resultando no aumento do tempo de resposta da rotina que desempenha um papel crítico para toda a aplicação.

Este problema afeta a integridade e confiabilidade de aplicações de jogos para servidores, comprometendo o seu devido funcionamento sem a necessidade de alocação de melhores recursos computacionais, apenas trabalhando com os já existentes.

A necessidade de alocação e investimento de máquinas mais robustas para o devido funcionamento da aplicação de jogos em servidores, bem como alto custo para manutenção e mitigação deste problema abordando apenas o paradigma OOP, onde é necessário o contrato de profissionais mais experientes para apoio técnico de *tunning* ou utilização de outras tecnologias para serem acopladas na aplicação e assim possibilitar uma falsa sensação de solução para este problema.

Solução Proposta

Utilizar os conceitos e fundamentos multidisciplinares do curso de graduação de Engenharia de Software com o padrão de desenvolvimento DOD, para auxiliar em uma melhor compreensão deste problema sem a necessidade de alto investimento financeiro para migração de outras tecnologias, pois este estudo tem como base providenciar uma sugestão de solução para este problema específico nos quais grandes corporações recentemente utilizaram para outros fins, porém de mesma complexidade e necessidade.

Sendo assim, proponho a devida interação multidisciplinar para fundamentar os conceitos que serão relacionados no decorrer deste projeto.

O que se esperar

Ao final deste projeto, é esperado alcançar um resultado satisfatório provenientes de testes, citações científicas, referências bibliográficas acadêmicas, bem como aplicação dos fundamentos multidisciplinares citados na Contextualização deste DVE em prática.

Objetivo Geral

O objetivo deste documento é contribuir para o aprimoramento e aperfeiçoar a qualidade das informações ofertadas, provenientes de pesquisas e estudos científicos. Sendo assim, este projeto tem como objetivo geral englobar várias áreas de conhecimento que estão interligadas de modo a atender as necessidades essenciais para o devido atendimento destes requisitos propostos na Solução Proposta deste DVE.

Objetivos Específicos

Utilizando como referência principal, mas não somente analisando de forma genérica certos fundamentos e conceitos teóricos, este projeto tem como ênfase específica nos seguintes objetivos:

I. Compreender o devido funcionamento de procedimentos de encapsulamento da mensagem através do paradigma OOP dentro do ambiente computacional;

II. Analisar de forma conceitual partições da memória gerenciadas pelo sistema operacional de modo a atender as necessidades de compreensão para o padrão de desenvolvimento DOD;
III. Interpretar as vantagens e desvantagens entre o paradigma OOP e o padrão de desenvolvimento DOD;
IV. Proporcionar através de uma nova perspectiva a devida relação entre o padrão de desenvolvimento DOD com recursos computacionais, sendo estes o processador e procedimentos de alocação de memória;
V. Capacitar possibilidade de generalização do padrão de desenvolvimento DOD para outros fins que atendem as mesmas características publicadas através dos relatórios de teste e demandam mesma aplicabilidade.

Estrutura do Projeto

Este projeto foi alocado nos seguintes capítulos de modo a auxiliar o seu devido entendimento e compreensão com a linha de pesquisa adotada, sendo estes:

- Capítulo 1 – alocação de memória durante procedimentos instanciação de objetos;
- Capítulo 2 – quando o paradigma OOP se torna um problema;
- Capítulo 3 – funcionamento da memória dentro do ambiente computacional;
- Capítulo 4 – operações assíncronas para otimização de processos através dos fundamentos *multithreading*;
- Capítulo 5 – OOP *versus* DOD;
- Capítulo 6 – os motivos por trás do DOD;
- Capítulo 7 – a devida utilização DOD em operações de alta performance e baixa latência de resposta em aplicações *server-side* de jogos;
- Capítulo 8 – as aplicações do padrão de desenvolvimento DOD pela empresa Unity em sua nova tecnologia, o Unity DOTS;
- Conclusão;
- Referências Bibliográficas;
- Glossário;
- Anexos;
- Apêndices.

Capítulo 1 – alocação de memória durante procedimento de instanciação de objetos

Quando trabalhamos com procedimentos do paradigma de orientação à objetos ou OOP, devemos ter em mente como de fato as ações de instanciação são organizadas na memória. Para isso é necessário compreender que existem duas maneiras[1] de alocar dados na memória, como:

- **Alocação de memória através da compilação:** caracterizado ou não por alocação estática, pois a memória requerida que deverá ser alocada por variáveis, tamanhos fixos, tipos de dados a serem trabalhados e conjunto padrão de coleções e definições devem ser conhecidos e alocados neste procedimento durante a compilação do programa; e
- **Alocação de memória dinâmica:** este procedimento é caracterizado por uma alocação dinâmica da memória, ou seja, no decorrer do uso da aplicação a memória exigida é devidamente realocada de acordo com os segmentos programados e definições de tamanho ou quantidade de espaço exigido que não precisam ser conhecidas pelo compilador. Entretanto, apesar de criar vários itens ou instâncias de objetos que serão alocados na memória, a alocação dinâmica de memória não fornece suporte para criar novos nomes de itens, ou seja, através de operações programadas utilizam-se recursos já existentes naquelas operações para assim criar através de forma dinâmica um espaço na memória para aquele novo item que baseia-se nas definições pré-definidas do algoritmo.

Todavia, durante a instanciação de objetos na memória deve ser realizado com cautela. Tanto em ações de criação quanto em destruição, é necessário ter essa preocupação do programador, pois a alocação exacerbada de novas instâncias de objetos na memória sem o devido tratamento de procedimentos de remoção dos espaços alocados, caracteriza uma situação de **vazamento de memória** ou *memory leak*[2]. Entretanto, é importante ressaltar que somente classes de objetos que serão instanciadas na memória de forma dinâmica, ou *constructor*, necessitarão de procedimentos de liberação dos espaços alocados, conhecido pela terminologia *deconstructor*.

Existem linguagens e *frameworks*, que provém a liberação de espaços alocados da memória de forma a diminuir essa preocupação do programador, como por exemplo: nas linguagens Java e C# (.NET) a liberação de instâncias alocadas de objetos sem a devida utilização são constantemente descartadas através do ***garbage collector***[2] e suas gerações que operam como auxiliadores na coleta de lixo para evitar estouros de memória durante a operação da aplicação no sistema operacional.

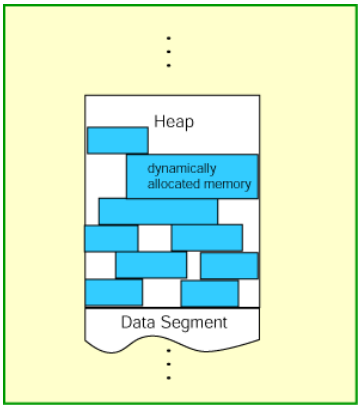

Figura 1.1 – caracterização de buracos ou espaços na memória através do problema de *memory fragmentation*. Fonte: Central Connecticut State University.

Basicamente, a utilização da alocação de memória dinâmica parte do princípio na OOP de que a memória é fornecida como um facilitador para auxiliar as requisições de retorno de endereços onde os dados podem ser pesquisados, verificados e alterados de forma dinâmica, assim facilitando os procedimentos para essas operações complexas. Porém, situações como criação de vários buracos ou *gaps* dentro da memória através de ações de criação ou destruição, podem desenvolver problemas que fazem deste conceito tornar-se ineficiente, pois essa situação é conhecida como **fragmentação da memória** ou *memory fragmentation*[3], como pode ser analisado na **figura 1.1** os espaços criados. Felizmente, essa situação pode ser mitigada em sistemas operacionais modernos.

Capítulo 2 – quando o paradigma OOP se torna um problema

O paradigma de orientação à objetos é ainda muito utilizado no desenvolvimento de software atualmente e consigo trás várias abordagens, nos quais auxiliam o programador a efetuar operações de diferente nível de complexidade com a devida compreensão deste paradigma. A implementação deste tipo de paradigma fornece a capacidade ao programador em compreender procedimentos de engenharia de uma perspectiva mais simples e fácil.

Algumas das principais vantagens quanto a utilização do paradigma OOP, podem ser listadas abaixo:

- **Herança e virtualização de métodos[6][7]:** permite que objetos sejam definidos de modo a permitir que todos seus dados e métodos existentes ao seu antecessor sejam herdados. Entretanto, alguns métodos oriundos da definição de métodos antecessores possuem a capacidade de virtualização, ou seja, apesar de herdarem os comportamentos pré-definidos, estes são capazes de comportarem-se de forma diferente e única;
- **Encapsulamento[8]:** aplicação estrita da ocultação da informação. Esta técnica tem como minimizar a dependência através da separação de objetos definidos através de interfaces externas, ou seja, auxilia a reduzir o acoplamento e a implementação de abstração de tipos de dados, de modo a promover e melhorar a manipulação destes parâmetros pertencentes ao objeto. Esta vantagem fornece suporte ao programador de proteger os dados internos daquele objeto de modo a promover a devida utilização, e funcionalidade correta, das propriedades, de modo a evitar retrabalho e possíveis ocorrências indesejadas através de atribuições não validadas para aquele objeto;
- **Reutilização[8]:** habilidade de reutilizar o sistema, ou partes dele, na construção de novos sistemas. Apesar de não ser algo pertencente aos componentes do paradigma OOP, esta característica pode ser encontrada em procedimentos onde estes componentes específicos, como classes, possam ser combinados e modificados para atender as devidas demandas de uma nova aplicação e para isso este paradigma pode ser essencial quando combinado a elementos de polimorfismo e herança; e
- **Extensibilidade[8]:** facilidade no qual o software do sistema tem de ser alterado ou modificado. Bem como a reutilização do software, ou de partes dele, em aplicações de outros sistemas corroboram em outra característica oriunda deste paradigma.

Logo, é notável as vantagens quanto o uso do paradigma de orientação à objetos e os futuros benefícios quanto utilização e aplicação no desenvolvimento de softwares. Todavia, nesta pesquisa venho trazer uma outra concepção e abordagem em que, estas características do OOP podem ser prejudiciais para manipulação de dados em procedimentos que exigem processamento com tempo de latência baixo devido aos conceitos de encapsulamento e organização dos dados na memória.

Capítulo 3 – funcionamento da memória dentro do ambiente computacional

Para abordar sobre o funcionamento da memória no ambiente computacional é necessário compreender alguns conceitos, como por exemplo *cache*. Desde os anos de 1980, o desenvolvimento dos microprocessadores recebeu inúmeras melhorias para melhor desempenho dentro do ambiente computacional. Entretanto, a evolução dos componentes para aumento de performance no acesso e uso da memória não cresceu na mesma proporção e isto proporcionou um buraco entre essas tecnologias, como pode ser observado na **figura 3.1**[9].

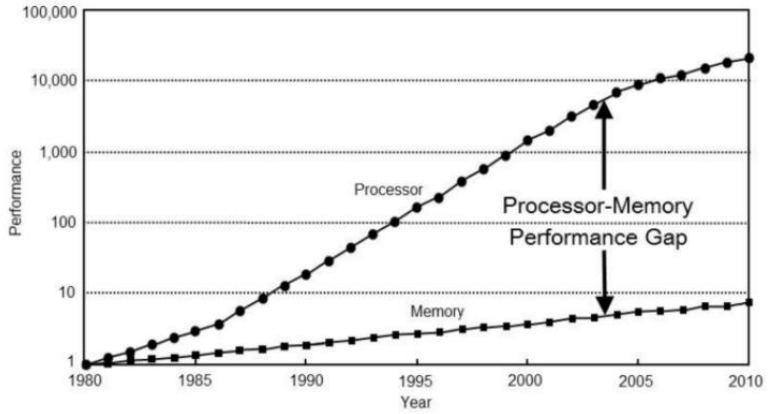

Figura 3.1 – comparativo de performance entre as tecnologias utilizadas nos processadores e memórias para ambientes computacionais. Fonte: BitSquid.

Este cenário motivou que grupos de pesquisas proporcionassem uma solução para este problema. Sendo assim, o conceito de *cache* é implementado no microprocessador, onde pequenas porções de memória são utilizadas para armazenar informações geridas pela CPU e que serão constantemente requisitadas para agilizar procedimentos de consulta a estes dados temporariamente salvos[10].

Agora, com essa nova abordagem, microprocessadores poderiam proporcionar um suprimento direto para acesso rápido da memória evitando acessar outros níveis, no qual proporciona maior tempo de busca e obtenção das informações contidas na memória. O funcionamento da memória é subdividido em níveis ou *layers*, nos quais cada nível é responsável em guardar dados que serão utilizados durante execução de softwares. Quanto maior o nível, mais tempo de latência ou maior o número de *clocks* serão exigidos para o processador para obter ou alterar aquele dado e esta característica de não encontrar uma informação na memória é conhecida como ***cache miss***[11] que pode comprometer a performance da aplicação. Um exemplo de uma operação de *cache miss*, pode ser demonstrada através da **figura 3.2** onde uma aplicação não encontra um dado dentro do *cache* e requisita uma consulta à uma base de dados.

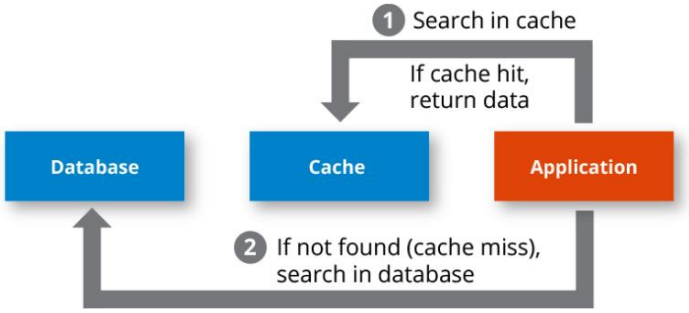

Figura 3.2 – demonstração de uma operação *cache miss* em um ambiente computacional por uma aplicação através de verificação *cache* e requisição à uma base de dados. Fonte: Hazelcast.

O processo unitário para adicionar mais *caches* dentro de um microprocessador pode ter consequências ou penalidades de tempo de resposta, com citado anteriormente através de vários ciclos ou *clocks*. Como mencionado pela patente publicada pela IBM em 2000[12], o *cache* L2 atua como intermediador entre o sistema de memória primário e os componentes à bordo de *caches*, de modo a proporcionar um espaço maior no armazenamento de dados ou instruções que os *caches* padrões, no qual pode ser observada na **figura 3.3**.

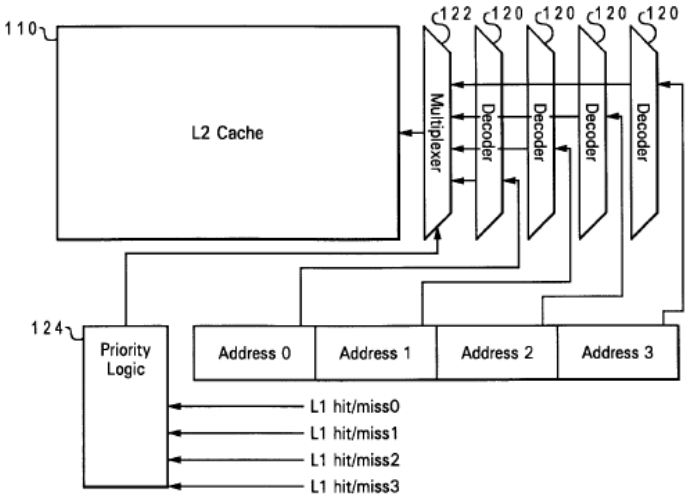

Figura 3.3 – diagrama representativo do *caches* L1 e L2 para acesso de memória sobreposta a múltiplos níveis dentro de uma arquitetura de microprocessador. Fonte: IBM Corporation.

Capítulo 4 – operações assíncronas para otimização de processos através dos fundamentos *multithreading*

Com a evolução dos sistemas operacionais foram possibilitadas operações nas quais auxiliavam a solucionar de produtividade nestes ambientes, como para uso científico ou comercial quanto ao uso adequado do processador[5]. Alguns problemas que eram evidenciados naquela época foram *CPU Bounded* e *I/O Bounded*, onde estavam relacionados a taxa de processamento e cálculo para instruções e tráfego de entrada-saída de dados através de periféricos pertencentes ao computador.

Contudo, os sistemas operacionais possibilitaram que mecanismos de sistemas multiprogramação junto do conceito de pseudoparalelismo permitissem que procedimentos de múltiplos processos

fossem compartilhados simultaneamente. Essa concepção possibilitou o aumento da produtividade de acordo com o crescimento de quantidade de tarefas completas por unidade de tempo, também conhecida como a característica ***throughput***[5].

Para ser abordado procedimentos de *multithreading*, é necessário compreender que esta ferramenta está comumente presente em diversos *frameworks*, como por exemplo .NET Framework 4[13]. Como é citado pela API de referência da Microsoft sobre *multithreading*, a utilização da programação *multithread* permite que a aplicação promova uma interação de melhor qualidade ao usuário que a utiliza.

Alguns procedimentos de multitarefas ou *multithreading*, podem ser caracterizados em operações paralelas, nos quais dividem-se em processos paralelos[5]:

- **Independentes:** são processos paralelos que quando utilizam recursos completamente distintos, não se relacionam em disputas com outros processos;
- **Concorrentes:** são processos paralelos que quando pretendem utilizar um mesmo recursos, dependem de uma ação do sistema operacional para definir a ordem de utilização; e
- **Cooperantes:** é quando dois ou mais processos utilizam em conjunto, um mesmo recursos para completarem uma dada tarefa.

Logo, esta atribuição fornece suporte que processos hajam de forma independente com as características especificadas anteriormente, onde pode ser observado na **figura 4.1**[4]. Todavia, a utilização de procedimentos de multitarefas não é necessariamente um fator de aumento de performance, pois com o uso indevido ou em excesso pode proporcionar uma queda de qualidade no tempo de resposta de uma aplicação[4].

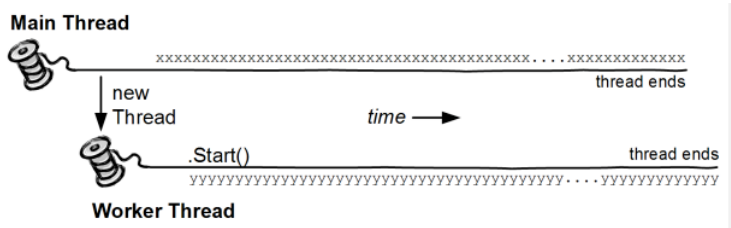

Figura 4.1 – representação de um *thread* em execução independente. Fonte: O'Reilly Media, Inc.

Sendo assim, em meio ao ambiente computacional para desenvolvimento de aplicações *server-side* de jogos, foi crucial a implementação destes fundamentos para proporcionar aumento de performance em procedimentos, processos ou tarefas que se caracterizavam como exaustivos para a plataforma computacional de modo a comprometer a integridade do ambiente ou aplicação.

Capítulo 5 – OOP *versus* DOD

Texto capítulo 5.

Capítulo 6 – os motivos por trás do DOD

Texto capítulo 6.

Capítulo 7 – a devida utilização DOD em operações de alta performance e baixa latência de resposta em aplicações *server-side* de jogos

Texto capítulo 7.

Capítulo 8 – as aplicações do padrão de desenvolvimento DOD pela empresa Unity em sua nova tecnologia, o Unity DOTS

Texto capítulo 8.

Conclusão

Texto conclusão.

Referências Bibliográficas

[1]. “Dynamic Memory Allocation”, Department of Computer Science, Florida State University, disponível em https://www.cs.fsu.edu/~myers/c++/notes/dma.html ;
[2]. “Memory allocation and deallocation (in relation to OOP)”, University of Washington, disponível em http://courses.washington.edu/css342/zander/css332/memoryoop.html ;
[3]. “Dynamic Memory Allocation”, Central Connecticut State University, disponível em https://chortle.ccsu.edu/assemblytutorial/chapter-33/ass33_3.html#:~:text=The%20program%20then%20uses%20this,part%20of%20the%20data%20segment. ;
[4]. “Threading in C#”, Albahari, J., O’Reilly Media, Inc., disponível em http://www.albahari.com/threading/ ;
[5]. “Notas sobre Sistemas Operacionais – Sistemas operacionais: processamento de dados”. Jandl, P. Jr., fev. 2004, versão 1.1;
[6]. “Object Oriented Programming of the finite element method”, R.I. Mackie, 1992, Elsevier Science Publishers Ltd. ;
[7]. “Object Oriented Programming and Numerical Methods”, R.I. Mackie, 2008, John Wiley & Sons, Ltd.;
[8]. “Encapsulation, Reusability and Extensibility in Object-Oriented Programming Languages”, M. Josephine, ago. 1987, Department of Computer Science, Columbia University ;
[9]. “Practical Examples in Data Oriented Design”, Frykholm, N., BitSquid, disponível em https://docs.google.com/presentation/d/17Bzle0w6jz-1ndabrvC5MXUIQ5jme0M8xBF71oz-0Js/present?slide=id.i16 ;
[10]. “How L1 and L2 CPU Caches Work, and Why They’re an Essential Part of Modern Chips”, Hruska, J., Extreme Tech, disponível em https://www.extremetech.com/extreme/188776-how-l1-and-l2-cpu-caches-work-and-why-theyre-an-essential-part-of-modern-chips ;
[11]. “What Is a Cache Miss?”, Hazelcast, Inc. disponível em https://hazelcast.com/glossary/cache-miss/ ;
[12]. “Multiple level cache memory with overlapped L1 and L2 memory access”, Dhong, S. H.; Hofstee, H. P.; Meltzer, D.; Falls, W.; Sillberman, J. A., out. 2000, IBM Corporation ;
[13]. “Managed Threading”, Microsoft, disponível em https://docs.microsoft.com/en-us/dotnet/standard/threading/ ;
[14]. “Pitfalls of Object Oriented Programming”, Albrecht, T., Research & Development Division, Sony Computer Entertainment Europe ;

Glossário

Cache – é uma pequena porção da memória que pode armazenar informações ou dados ou instruções que serão utilizadas pela CPU[10].
Cache miss – é uma operação no qual uma informação não é encontrada naquele contexto e assim recorre aos próximos níveis de memória, proporcionando perda de performance na aplicação através de ciclos de processamento extra realizados pelo processador[10].
DOD – *Data-Oriented Design*, é um padrão de desenvolvimento para uma abordagem de otimização mais eficiente do uso de CPU mediante o desenvolvimento de jogos. Fonte: "Data-oriented design", Wikipedia.
Multithreading – são ferramentas disponibilizadas através de *frameworks* que possibilitam a utilização dos recursos computacionais do ambiente, de modo a possibilitar processos de paralelismo para assim possibilitar aumento de performance na execução de aplicações[4].
OOP – *Object-Oriented Programming*, é um paradigma de orientação à objeto utilizado na programação baseado nos conceitos de objeto, nos quais estes podem ser acessados e conter dados ou informações e assim serem facilmente manipulados através de procedimentos, por exemplo polimorfismo, encapsulamento etc. Fonte: "Object-oriented programming", Wikipedia.
Processo computacional – é uma atividade que ocorre no meio computacional, usualmente possuindo um objetivo definido, tendo duração finita e utilizando uma quantidade limitada de recursos computacionais, onde é caracterizado pelo termo tarefa ou *task*[5].
Processos paralelos – são executados ao mesmo tempo, sendo caracterizados como independentes, concorrentes e cooperantes[5].
Processos paralelos concorrentes – são processo que desejam utilizar uma região crítica e assim dependem de uma ação do sistema operacional para definir a ordem de utilização[5].
Região crítica – *critical section*, ocorre quando um dado recurso computacional só pode ser utilizado por um único recurso simultaneamente[5].
Threads – são fluxos independentes de execução, onde pertencem a um mesmo processo, ou seja, requerem menos recursos de controle por parte do sistema operacional[5].

Anexos

Texto anexos.

Apêndices

Texto apêndices.