SSC-0158
Computação em Nuvem

Aula 08 – Escalonamento

Prof. Julio Cezar Estrella
jcezar@icmc.usp.br

Créditos

Os slides integrantes deste material foram construídos a partir dos conteúdos relacionados às referências bibliográficas descritas neste documento.

Introdução

- É uma atividade organizacional feita pelo escalonador (scheduler) da CPU ou de um sistema distribuído, possibilitando executar os processos mais viáveis e concorrentes, priorizando determinados tipos de processos, como os de I/O Bound e os CPU Bound.
- Escalonador de Processos escolhe o processo que será executado pela CPU;
- O escalonamento é realizado com o auxílio do hardware; O escalonador deve se preocupar com a eficiência da CPU, pois o chaveamento de processos é complexo e custoso:
  - Ele afeta desempenho do sistema e satisfação do usuário; O escalonador de processo é um processo que deve ser executado quando da mudança de contexto (troca de processo);

Escalonamento

- O escalonador do SO utiliza alguns critérios de escalonamento, como:
  - a taxa de utilização de CPU, que é a fração de tempo durante a qual ela está sendo ocupada;
  - throughput que são números de processos terminados por unidade de tempo;
  - turnaround que é o tempo transcorrido desde o momento em que o software entra e o instante em que termina sua execução;
  - tempo de resposta: intervalo entre a chegada ao sistema e inicio de sua execução;
  - tempo de espera: soma dos períodos em que o processo estava no seu estado pronto.

Escalonamento

- O projeto de uma política de escalonamento deve contemplar em geral os seguintes objetivos:
  - **Ser justo:** Todos os processos devem ser tratados igualmente, tendo possibilidades idênticas de uso do processador, devendo ser evitado o adiamento indefinido.
  - **Maximizar a produtividade** (throughput): Procurar maximizar o número de tarefas processadas por unidade de tempo.
  - **Ser previsível:** Uma tarefa deveria ser sempre executada com aproximadamente o mesmo tempo e custo computacional.
  - **Minimizar o tempo de resposta** para usuários interativos.
  - **Maximizar** o número possível de **usuário** interativos.
  - **Minimizar a sobrecarga** (overhead): Recursos não devem ser desperdiçados embora algum investimento em termos de recursos para o sistema pode permitir maior eficiência.

Escalonamento

- Há situações nas quais um processo de escalonamento é necessário
  - Quando um novo processo é criado;
  - Quando um processo terminou sua execução e um próximo processo pronto deve ser executado;
  - Quando um processo é bloqueado (semáforo, dependência de E/S), resultando que outro processo deve ser executado.
  - Quando uma interrupção de E/S ocorre, o escalonador deve optar por: executar o processo que estava esperando essa interrupção; continuar executando o processo que já estava sendo executado; ou executar um terceiro processo que esteja pronto para ser executado.

Algoritmos de Escalonamento

- A computação em nuvem propicia a ilusão de que os recursos computacionais são para o uso.
- Os usuários têm a expectativa de que a nuvem seja capaz de fornecer rapidamente recursos em qualquer quantidade e a qualquer momento.
- É esperado que os recursos adicionais possam ser providos, possivelmente de forma automática, quando ocorre o aumento da demanda e retidos, no caso da diminuição desta demanda.

Algoritmos de Escalonamento

- Preemptivos e Não Preemptivos
  - **Preemptivos:** são algoritmos que permitem que um processo seja interrompido durante sua execução, que seja por força de uma interrupção de entrada/saída, quer seja em decorrência da politica de escalonamento adotada e aplicada por parte do escalonador de processos ou simplesmente por força do término da execução do processo
  - Não Preemptivos: O processo executa até o fim, sem ser interrompido

Algoritmos de Escalonamento

- Exemplos de algoritmos
  - FIFO
  - SJF
  - SRT
  - RR
  - Múltiplas Filas
  - Fair-Share

Algoritmos de Escalonamento e Cloud

- Quando pensamos em cloud computing, disponibilidade, eficiência e economia são pontos que necessitam ser considerados
- O **escalonemto horizontal** é uma técnica de dimensionamento, que divide a carga de trabalho e o conjunto de dados do sistema nos servidores existentes ou ainda adicionando servidores extras para aumentar a capacidade conforme necessário.

Algoritmos de Escalonamento e Cloud

- Geralmente, é preciso apoio de um **load balancer**:
  - um componente responsável por distribuir solicitações de usuários (carga de trabalho) entre os vários sistemas, máquinas ou nós de back-end no cluster. Cada uma dessas unidades de back-end executa uma cópia de seu software e, portanto, é capaz de atender as solicitações com eficiência
  - No caso o **load-balancer** nada mais é que um escalonador que agirá segundo uma política de escalonamento
    - Veja a importância das estratégias de escalonamento que já conhecemos em Sistemas Operacionais, e que podem ser amplamente utilizadas no contexto de Cloud

Algoritmos de Escalonamento e Cloud

- A importância do escalonamento horizontal
  - Se o objetivo é executar aplicações em um volume cada vez maior, é interessante ter como referência o escalonamento horizontal na nuvem desde o início, como parte do processo de planejamento
    - Como essa norma de dimensionamento não pode ser implementada a qualquer momento, é preciso estruturá-la na arquitetura original

Algoritmos de Escalonamento e Cloud

- A empresas que fornecem serviços da Web, como Google, Microsoft, Facebook e Amazon usam intensamente o escalonamento horizontal. Entre os benefícios de adotar essa estratégia estão:
  - Alta disponibilidade do servidor;
  - Melhor distribuição da carga de trabalho nos nós existentes;
  - Rápida configuração de escalabilidade;
  - Automatização do desempenho;
  - Previsibilidade de custos e da carga de trabalho.

Algoritmos de Escalonamento e Cloud

- **Escalonamento Horizontal x Escalonamento Vertical**
  - O **escalonamento horizontal** envolve a adição ou remoção de unidades de processamento ou máquinas ao seu ambiente, conforme a sua necessidade instantânea. Isso vai aumentar ou diminuir o número de nós no cluster, redistribuindo a carga de trabalho entre os disponíveis. Esta ferramenta de dimensionamento tem sido amplamente usada para aplicações que exigem alto desempenho e controle de custos.
  - No **escalonamento vertical**, a diferença notável geralmente se dá pela limitação à capacidade de uma única máquina pois, para dimensionar sua aplicação além da capacidade do servidor, é preciso aumentar o hardware. No entanto, esse processo resulta em inatividade e custos mais altos.

Algoritmos de Escalonamento e Cloud

- **Escalonamento Vertical**
  - Para o aumento dos recursos verticalmente, é preciso aumentar a capacidade deste recurso. Por exemplo, se uma máquina virtual opera com 4 GB de memória RAM, um aumento vertical significaria que esta capacidade teria que ser aumentada para dar conta das demandas. Este tipo de dimensionamento costuma não ser utilizado, já que requer que o sistema fique temporariamente indisponível para a implementação do recurso.

Algoritmos de Escalonamento e Cloud

- **Escalonamento Horizontal**
  - Aqui a capacidade do recurso não é aumentada. Ela é duplicada, com as mesmas características da máquina virtual original. Se a máquina virtual original tem capacidades de memória com número x, as máquinas duplicadas terão a mesma capacidade. Essa é a forma mais utilizada porque permite que os recursos sejam provisionados sem que o sistema fique indisponível. Da mesma forma, caso a demanda caia, os recursos são desalocados automaticamente.
    - **Exemplo:** você tem uma estrutura padrão para um site de e-commerce, que demanda dois servidores virtuais em condições normais. Mas, por conta de uma campanha promocional, o acesso ao serviço aumentou dez vezes por dois dias.

Dúvidas