

**Projeto em ciência de
dados com soluções
para processamento
paralelo e distribuído
de dados**



Introdução a soluções para processamento paralelo e distribuído de dados

Bloco 1

Marcelo Tavares de Lima






► Objetivos

- Apresentar conceitos introdutórios de processamento paralelo de dados.
- Apresentar conceitos introdutórios de processamento distribuído de dados.
- Apresentar soluções para processamento paralelo e distribuído de dados e gerenciamento de *clusters* e *grids*.




► Introdução

- Introdução a soluções para processamento paralelo e distribuído de dados.
 - Gerenciamento de *clusters* e *grids*.
 - Conceitos fundamentais.
 - Diferenciais e outros.
- 




► Introdução

- Motivação principal: processamento de grandes bases de dados (*Big Data*).
 - Os sistemas precisam suportar o armazenamento e a execução.
 - Os sistemas precisam ser rápidos e ágeis.
 - Surgiram os sistemas paralelos e distribuídos de dados.
- 




► **Processamento paralelo**

- Surgiu com a intenção de redução do tempo de processamento de dados.
 - Surgiu na década de 90.
 - Quanto menor o tempo de resolução dos problemas, mais rapidamente se toma decisões importantes para os negócios.
 - O processamento também precisa ser confiável.
- 



► Processamento paralelo

- Na prática, é o uso de mais de uma unidade de processamento (CPU) para a execução da resolução conjunta de um problema.
 - Divide o problema em problemas (tarefas) menores.
 - Exige investimento em hardware e software.
 - Processamento de alto desempenho.
- 



► **Processamento paralelo**

Duas métricas avaliam a eficiência de um sistema em paralelo:

- Aceleração linear.
- Crescimento linear.



► Processamento paralelo

- Aceleração linear: avalia o tamanho do sistema.
- Exemplo: se o hardware for duplicado, uma tarefa poderá ser executada na metade do tempo utilizado se tivesse sido executada com um processador apenas.
- É medida como a razão entre o tempo de execução com um processador e o tempo de execução com mais de um processador.




► Processamento paralelo

- Crescimento linear: utilizada para medir a habilidade de crescimento do sistema e também do problema.
- Exemplo: se o hardware for duplicado, espera-se que o sistema passe a ser capaz de executar um problema, duas vezes mais, considerando o mesmo intervalo de tempo.



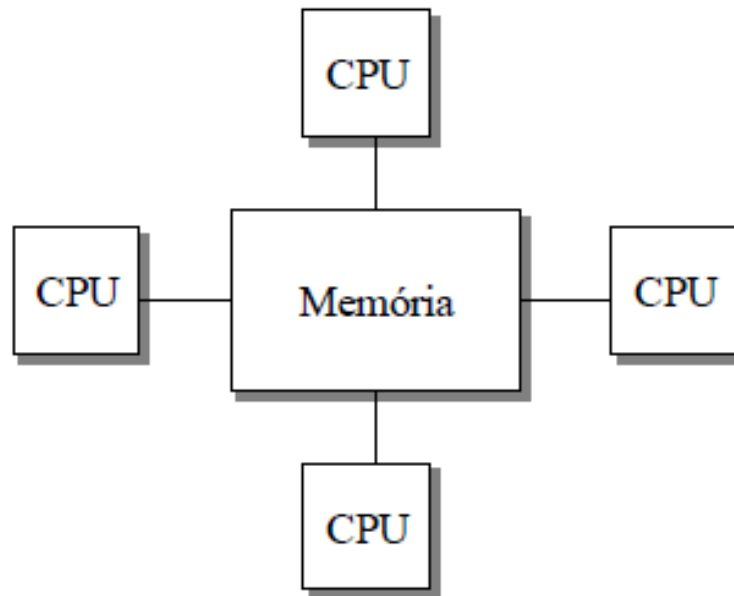
► **Processamento paralelo**

Vantagens do processamento paralelo, segundo Navaux, De Rose e Pilla (2011):

- Melhora no desempenho.
 - Maior tolerância a falhas.
 - Modela modelos mais complexos.
 - Aproveita mais os recursos.
- 

► Processamento paralelo

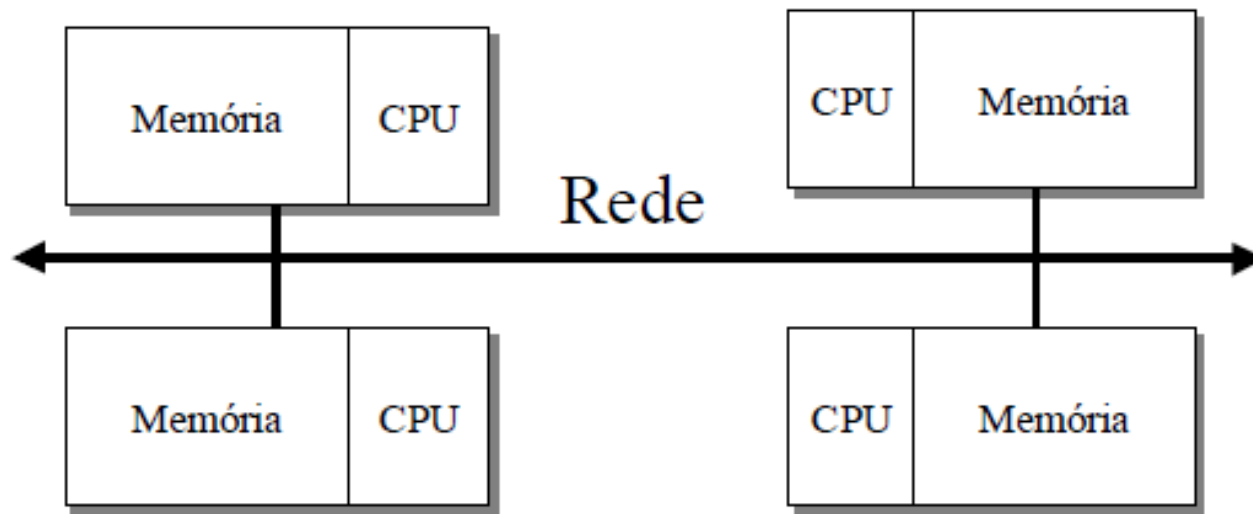
Figura 1 – Modelo de processamento paralelo: memória compartilhada



Fonte: Meyer (2006).

► Processamento paralelo

Figura 2 – Modelo de processamento paralelo: memória distribuída



Fonte: Meyer (2006).

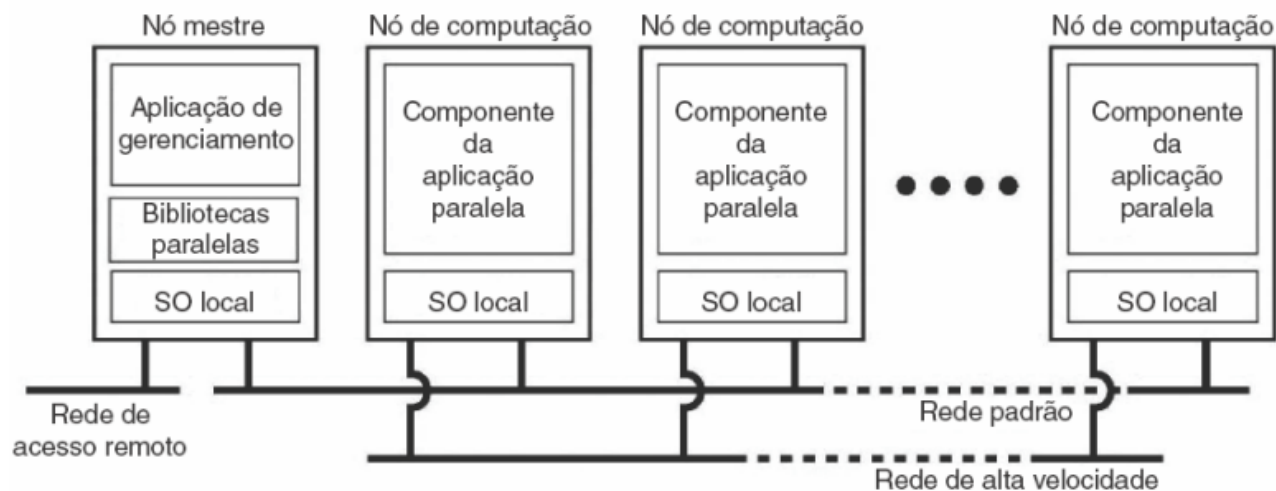


► *Clusters e grids*

Cluster é um sistema distribuído de computadores independentes e interligados, cujo o objetivo é suprir a necessidade de um grande poder computacional com um conjunto de computadores de forma transparente ao usuário. (BACELLAR, 2010, p. 3)

► Processamento paralelo

Figura 3 – Esquema de uma rede distribuída em *cluster*



Fonte: Martins (2019).



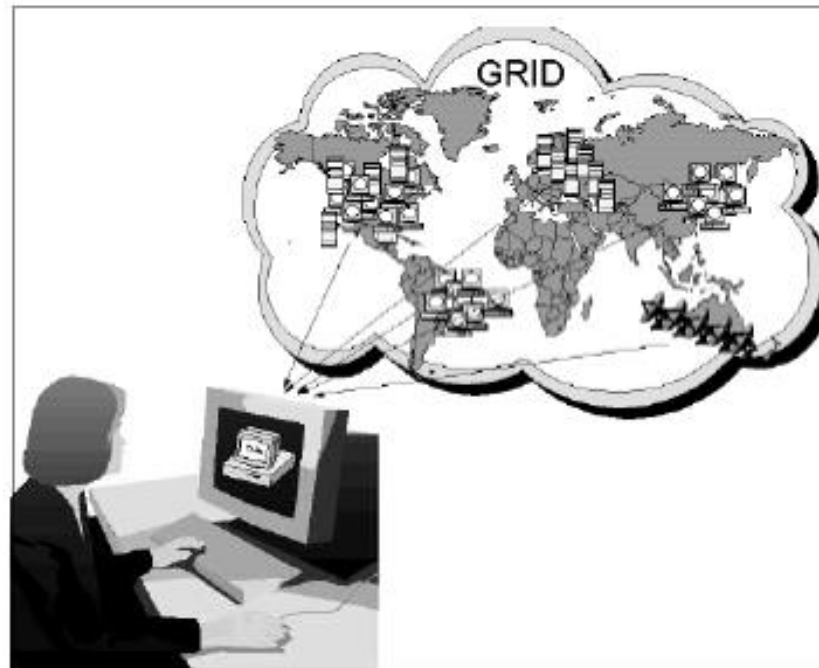
► *Clusters e grids*

Segundo Meyer (2006):

- A visão de *Grid* é similar (ou almeja ser) a uma rede de distribuição de energia elétrica, ou seja, diversos recursos computacionais geograficamente distribuídos podem ser agregados para formar um supercomputador virtual. (MEYER, 2006, p. 8)

► Processamento paralelo

Figura 4 – Modelo de processamento em *grid*



Fonte: Meyer (2006).

Introdução a soluções para processamento paralelo e distribuído de dados


Bloco 2

Marcelo Tavares de Lima





► Sistemas distribuídos

- É crescente o investimento em tecnologias, como os supercomputadores, em todas as áreas do conhecimento, em especial a ciência de dados e áreas especiais, como: a meteorologia (previsão do tempo); a busca por petróleo; área de simulações físicas; e a matemática computacional.
- 



► Sistemas distribuídos

- O processamento de alto desempenho é uma área da ciência da computação que veio para solucionar problemas complexos, como os encontrados nessas áreas, que antes eram resolvidos com a simplificação dos modelos, resultando em respostas menos precisas e com margem de erro considerável (NAVAUX; DE ROSE; PILLA, 2011).



► Sistemas distribuídos

- Muitos recursos utilizam computação paralela, como, por exemplo, aplicações em C ou em Java.
- As redes sociais que utilizamos hoje são exemplos de sistemas distribuídos.
- Sites de pesquisas e plataformas de vídeos on-line, como a Netflix, também são.



► Sistemas distribuídos

Quando se trabalha com sistemas distribuídos, existem os seguintes objetivos:

- Disponibilidade e acesso fácil ao sistema, assim como a todos os seus recursos, por todos os seus componentes, sejam máquinas ou usuários finais.
- Ocultar do usuário final que o sistema é distribuído.
- Facilitação da inclusão de novas máquinas, ou seja, deixar o sistema o mais aberto possível nesse sentido, para que possa expandir facilmente.

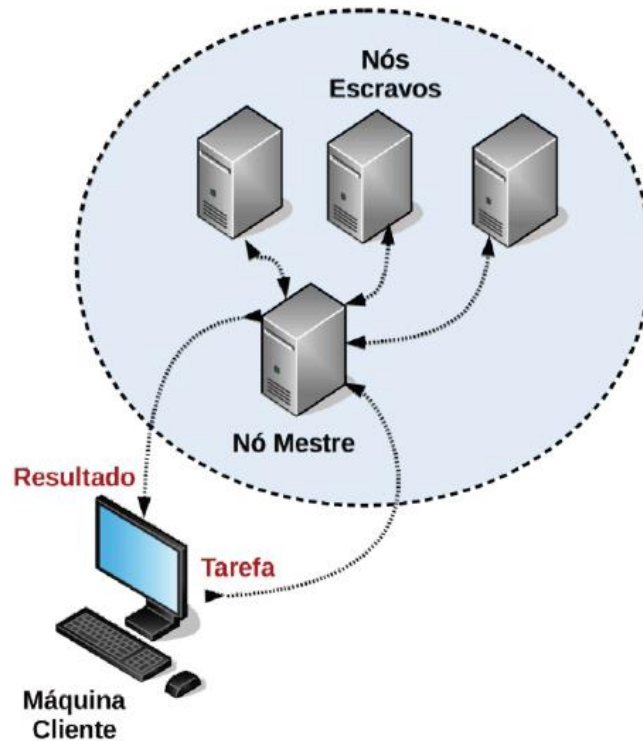


► Sistemas distribuídos

- Como já visto, os sistemas distribuídos podem ser classificados em *cluster* e *grid*.
- *Cluster*: conjunto de máquinas com hardwares semelhantes, características homogêneas interligadas por rede local (LAN).
- *Grid*: conjunto de máquinas com características diferentes. O hardware e os sistemas operacionais podem ser de fabricantes diversos.

► Sistemas distribuídos

Figura 5 – Sistema distribuído em *cluster*



Fonte: Pereira (2019).

► Sistemas distribuídos

Figura 6 – Sistema distribuído em *grid* - Cinegrid



Fonte: Pereira (2019).

PÓS-GRADUAÇÃO

Teoria em prática

Bloco 3

Marcelo Tavares de Lima





► Teoria em prática

- Imagine que você trabalha no departamento de pesquisa de mercado em uma empresa. Sua responsabilidade é gerenciar uma equipe de funcionários aptos para lidar com grandes bases de dados, do tipo *Big Data*.



► Teoria em prática

- Os equipamentos tecnológicos que sua equipe utiliza estão ficando obsoletos para lidar com bases de dados tão grandes quanto às que vocês conseguem manipular. Além disso, o sistema de gerenciamento de dados também está se tornando obsoleto.



► Teoria em prática

- A partir desse cenário, você reúne sua equipe e começa a planejar estratégias de melhorias do ferramental tecnológico e da rede de dados que fazem uso.



► Teoria em prática

- Uma rede com processamento paralelo resolve? Caso escolha esse tipo de sistema, é melhor um sistema de memória compartilhada ou distribuída?
- Entretanto, é possível concluir que é melhor um sistema de computação em *grids*, pois você precisa estar em rede com outras unidades da empresa.

Dica do professor

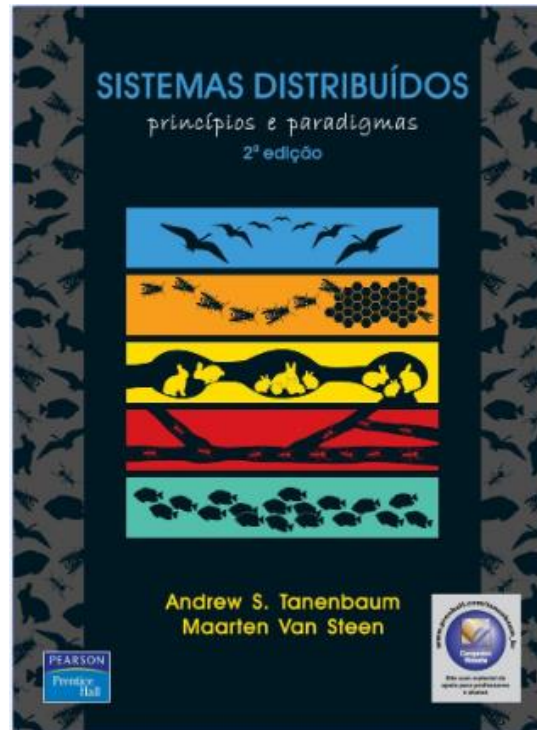
Bloco 4

Marcelo Tavares de Lima



► Indicação de livro

Figura 7 – Dica



Fonte:

https://bv4.digitalpages.com.br/?term=sistemas%2520distribu%25C3%25ADdos&searchpage=1&filtro=todos&from=busca&page=_1§ion=0#/legacy/411. Acesso em: 03 fev. 2020.



► Referências

BACELLAR, H. V. **Cluster**: computação de alto desempenho. Campinas: Instituto de Computação, Universidade Estadual de Campinas, 2010. Disponível em: <http://www.ic.unicamp.br/~ducatte/mo401/1s2010/T2/107077-t2.pdf>. Acesso em: 03 fev. 2020.

MARTINS, S. L. **Sistemas distribuídos**. Departamento de Ciência da Computação. Niterói: Universidade Federal Fluminense, 2019. Disponível em: ic.uff.br/~simone/sd/contaulas/aula2.pdf. Acesso em: 03 fev. 2020.

MEYER, L. A. V. C. **Uma visão geral dos sistemas distribuídos de cluster e grid e suas ferramentas para o processamento paralelo de dados**. 2006. IBGE [s.d.]. Disponível em https://ww2.ibge.gov.br/confest_e_confega/pesquisa_trabalhos/CD/palestras/368-1.pdf. Acesso em: 03 fev. 2020.



► Referências

NAVAUX, P. O. A.; de ROSE, C. A. F.; PILLA, L. L. **Fundamentos das arquiteturas para processamento paralelo e distribuído**. 2011. Laboratório de Banco de Dados. Departamento de Ciência da Computação – UFMG. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/erad-rs/2011/003.pdf>. Acesso em: 03 fev. 2020.

PEREIRA, C. S. **Sistemas distribuídos**. Londrina: Editora e Distribuidora Educacional S.A., 2019.

