Desafios de E/S em Ambientes de Grande Escala

Philippe O. A. Navaux
Francieli Zanon Boito
Rodrigo Virote Kassick

Grupo de Processamento Paralelo e Distribuído (GPPD)
Universidade Federal do Rio Grande do Sul (UFRGS)

ERAD-SP 2011 – São José dos Campos/SP

Roteiro

- Introdução
- Arquivo e Armazenamento
- Sistemas de Arquivos Paralelos
- Questões de Desempenho
- Tendências
- Conclusão

Roteiro

- **Introdução**
- Arquivo e Armazenamento
- Sistemas de Arquivos Paralelos
- Questões de Desempenho
- Tendências
- Conclusão

Introdução

- Desde o início da Informática houve a necessidade de **armazenar** as informações/ **dados** e os programas/ **instruções**, devido ao pouco espaço em memória.
- Para tal foram desnvolvidos mecanismos de armazenamento acessados por uma **entrada/saida E/S**:
  - **Fita magnética**
  - **Disco**
  - Outros meios.
- Hoje o principal meio de armazenamento secundário são os **discos**, que serão o objeto de nossa palestra.

Introdução

- **Operações de E/S mais lentas** que processamento
  - Dependem de discos, memória e interconexão
- Desempenho das aplicações é **limitado pelas operações de E/S**

Introdução

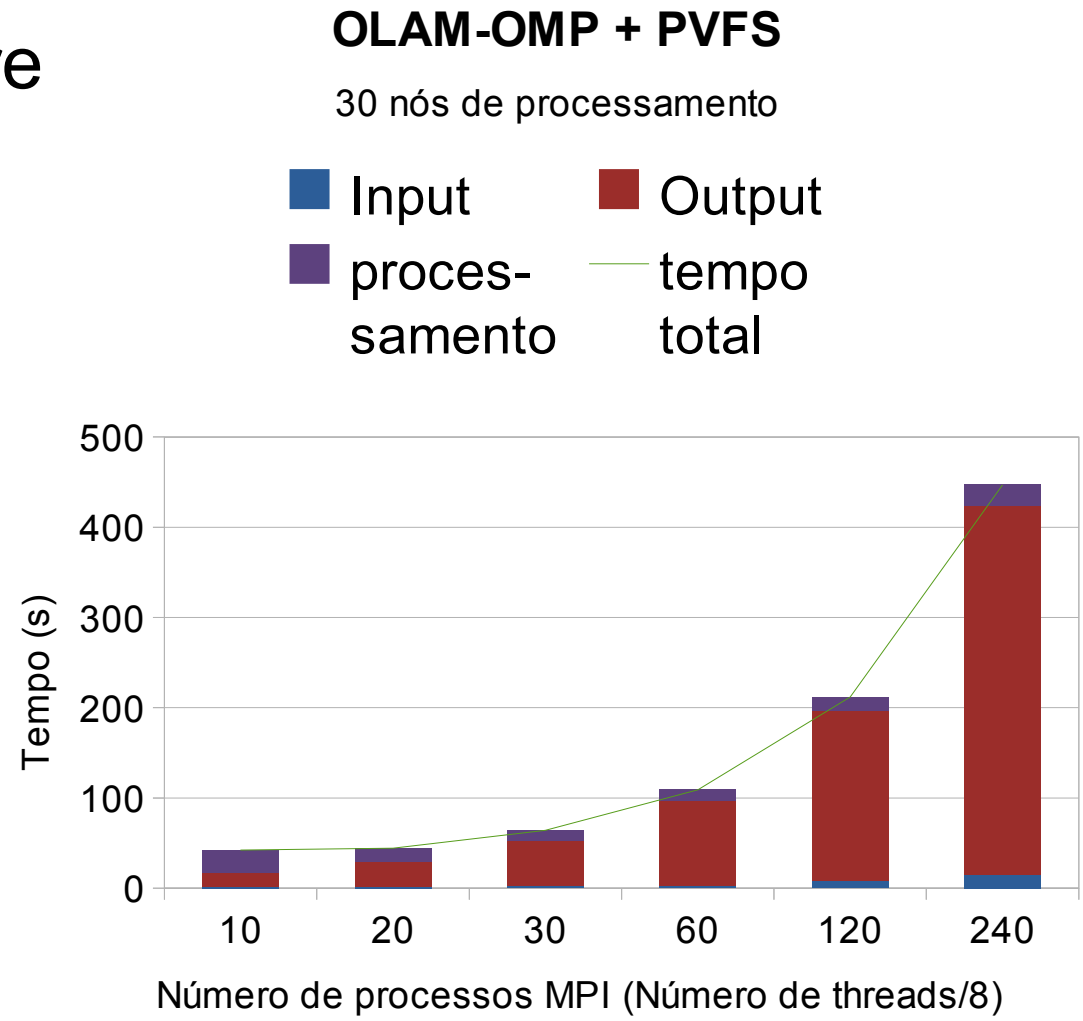
- Com o surgimento de sistemas com vários computadores/ processadores o problema somente aumentou.

- Aplicações executando no cluster precisam **compartilhar dados armazenados**
  - Ler dados de entrada, escrever resultados, checkpoint, …

- **Sistemas de Arquivos Paralelos SAP** permitem esse compartilhamento
  - Operações de E/S (ou I/O): acessos aos dados

Introdução

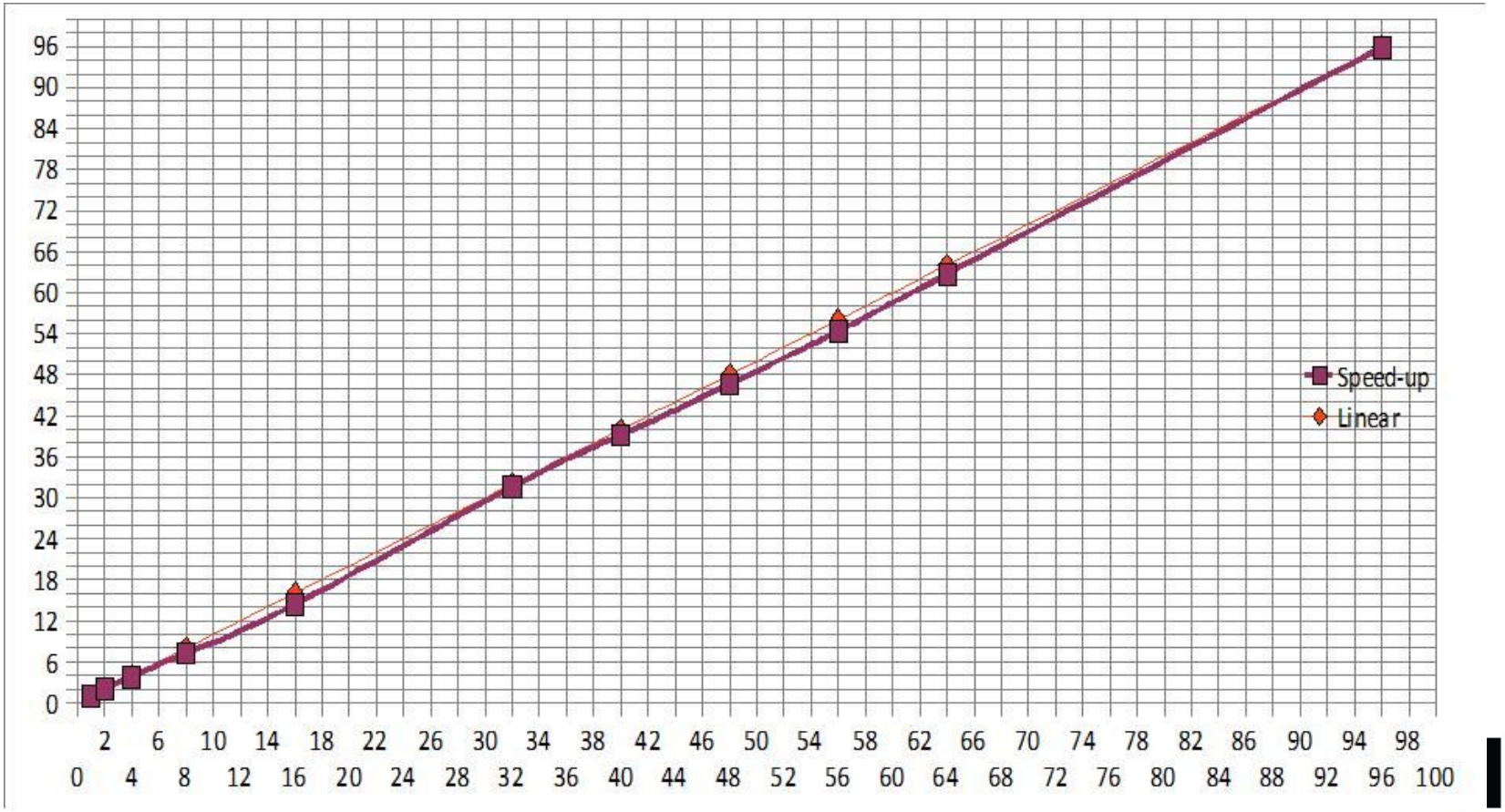
- Ocean-Land-Atmosphere Model (**OLAM**)
  - Modelo climático
  - Problemas de escalabilidade
  - Avaliação de [Boito et al. 2011]:



Introdução

OLAM - Avaliação de [Dias et al. 2010]:

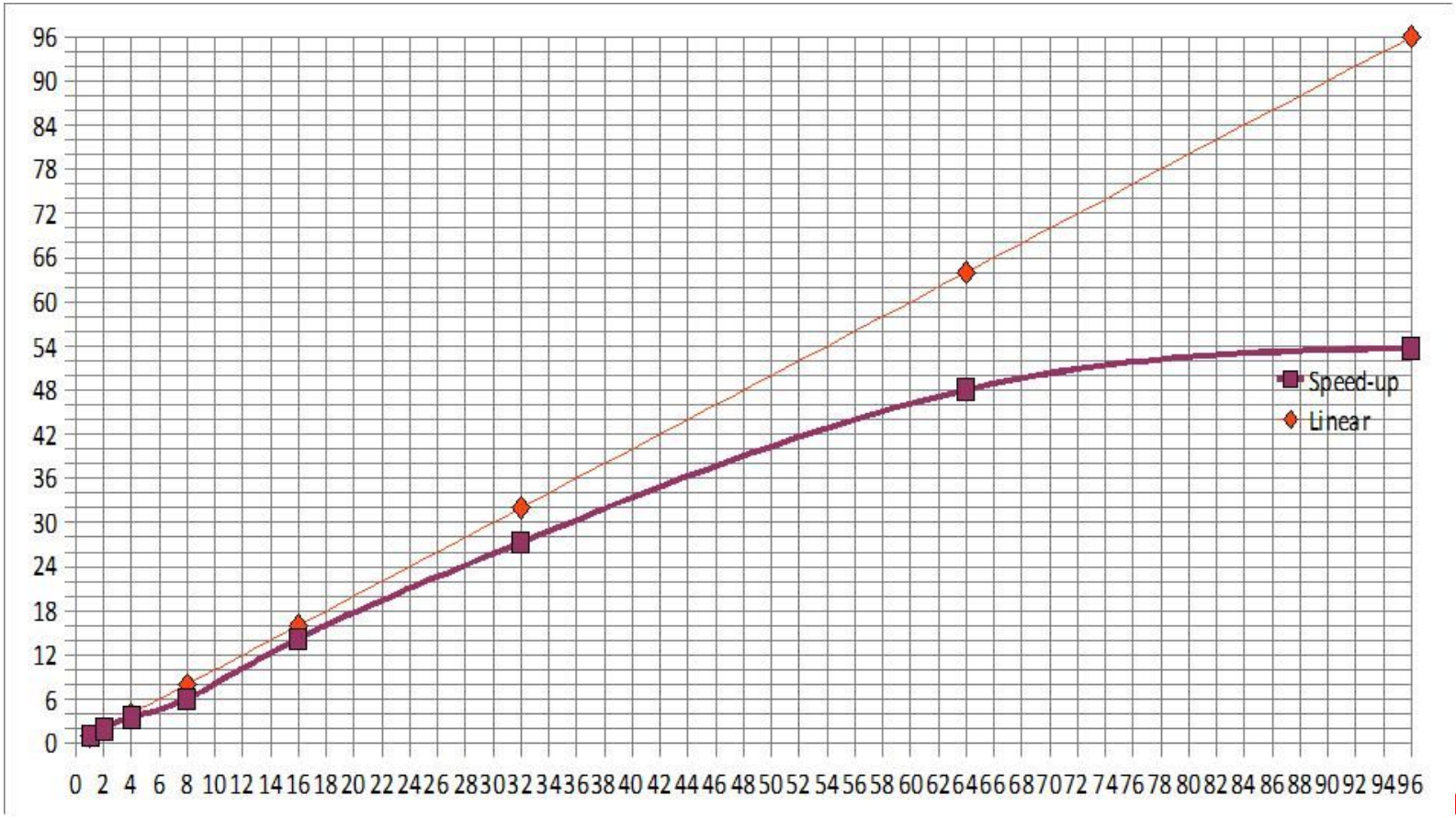
- Speed-up **SEM operações de E/S**



Introdução

OLAM - Avaliação de [Dias et al. 2010]:

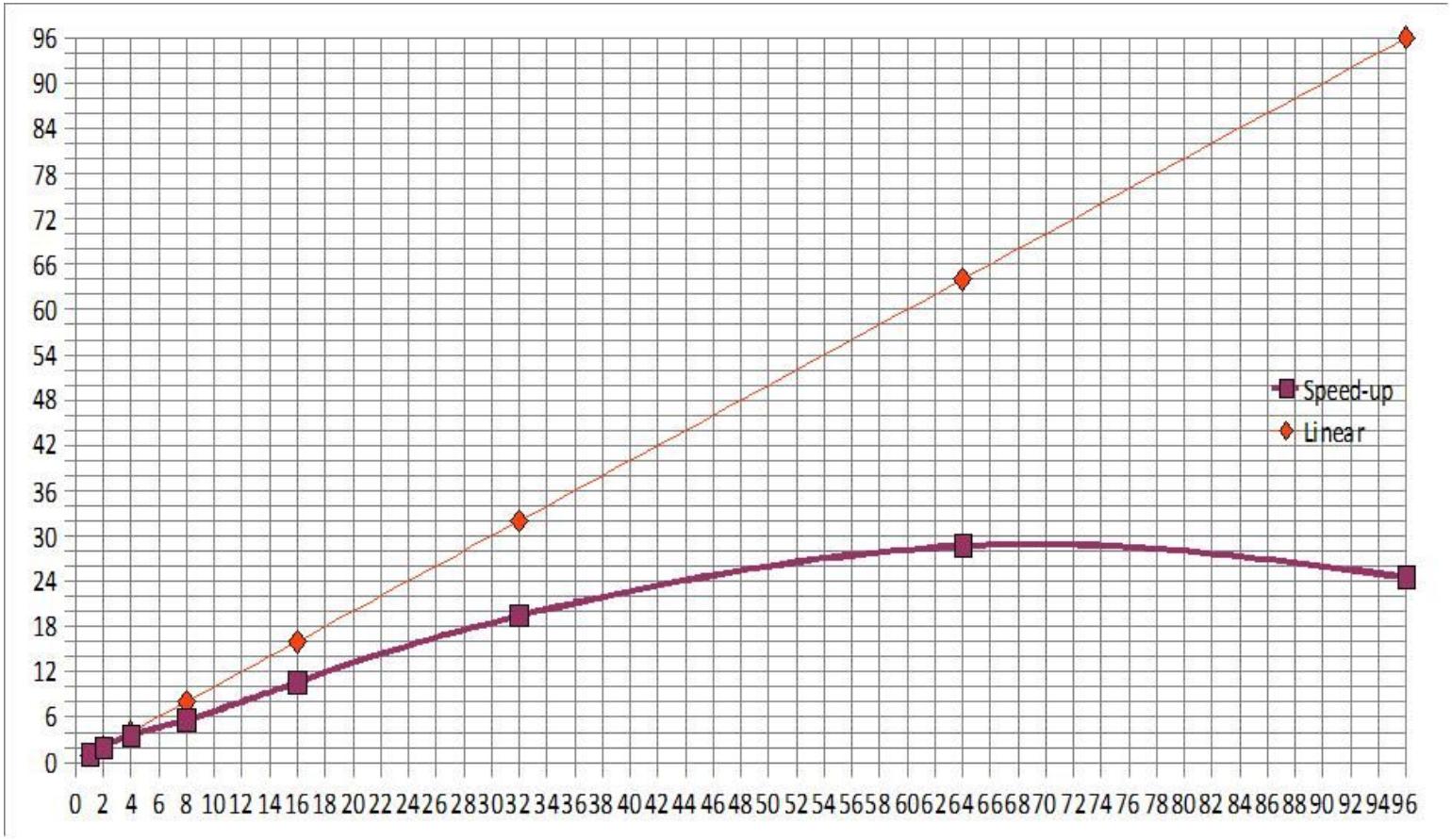
- Speed-up **SEM operações de E/S (apenas leitura)**



Introdução

OLAM - Avaliação de [Dias et al. 2010]:

- Speed-up **COM operações de E/S (leitura e escrita)**



Introdução

- Atualmente, temos supercomputadores **petascale**
  - 10^{15} operações de ponto flutuante por segundo (flops)

- **Próximo objetivo: exascale** (10^{18} flops)

"*performance cannot be scaled up by increasing the number of CPUs anymore, but by increasing the bandwidth of the I/O subsystem*" (Buyya, **1999**)

Introdução

- **I/O já é um gargalo** em sistemas atuais
- Problema será ainda mais sério em exascale!

Essa apresentação: mostrar **problemas e soluções** comuns da área de **E/S de alto desempenho**

Blue Waters


NCSA
IBM
BLUE WATERS
SUSTAINED PETASCALE COMPUTING
ILLINOIS

Roteiro

- Introdução
- **Arquivo e Armazenamento**
- Sistemas de Arquivos Paralelos
- Questões de Desempenho
- Tendências
- Conclusão

Arquivo e Armazenamento

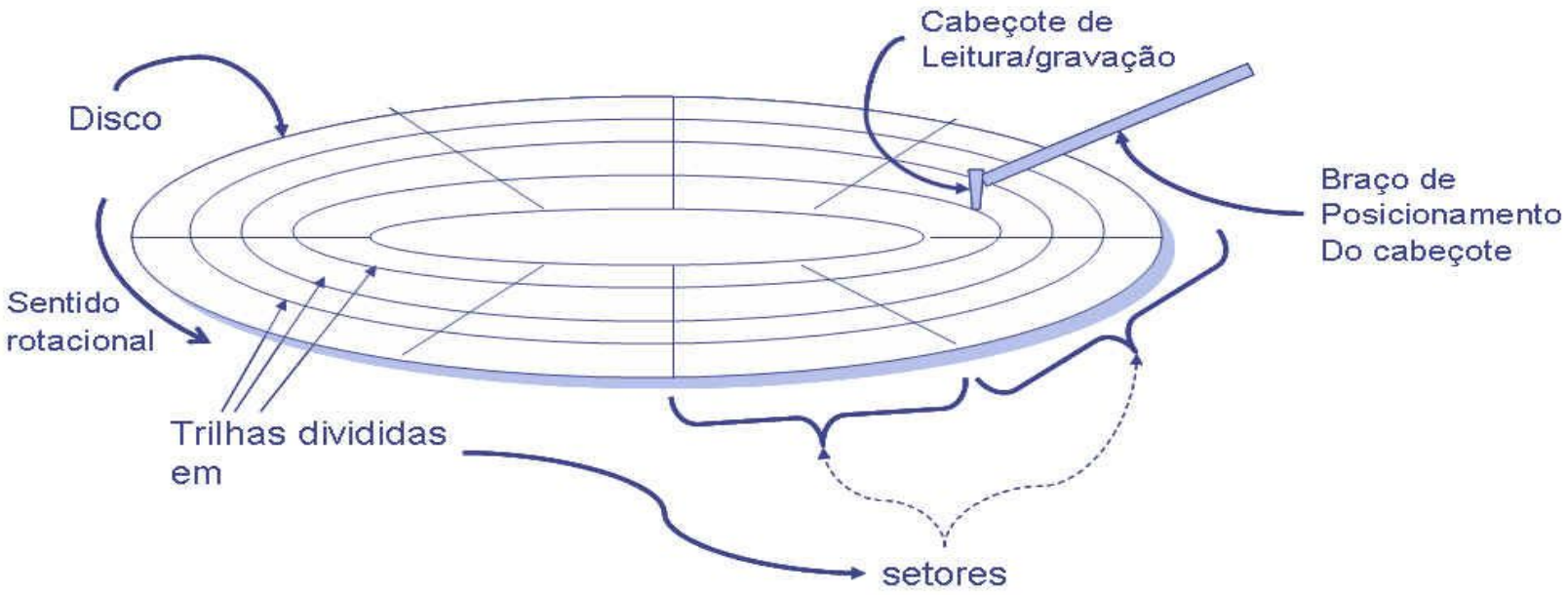
- **Arquivo** é uma abstração do Sistema Operacional
- Colocado no meio de **armazenamento persistente**
  - Discos => HDD, SSD, etc
- Armazenamento é separado em uma **série de blocos**
- Gerenciado pelo **Sistema de Arquivos**

Arquivo e Armazenamento

- **Sistema de arquivos** é responsável por:
  - Fornecer mecanismos para manipular os arquivos
  - Garantir coerência de dados
  - Otimizar o acesso
  - Suporte a vários usuários
  - ...

Arquivo e Armazenamento

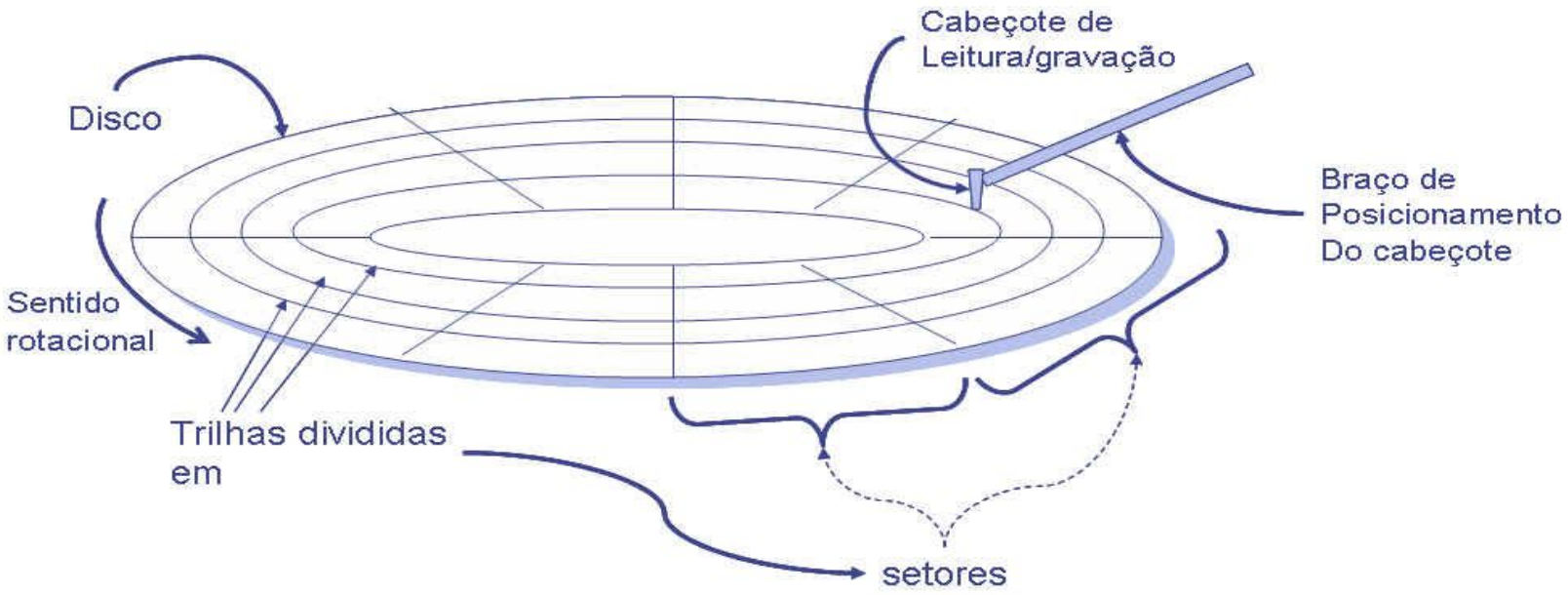
- **Hard Disk Drive (HDD)**
- Prato(s) circular(es) + cabeça de leitura/gravação



Arquivo e Armazenamento

- **Tempo para acesso:**

1. Tempo de busca (seek): mover a cabeça para a trilha certa
2. Latência rotacional: o início do bloco chegar à cabeça
3. Tempo de transferência



Arquivo e Armazenamento

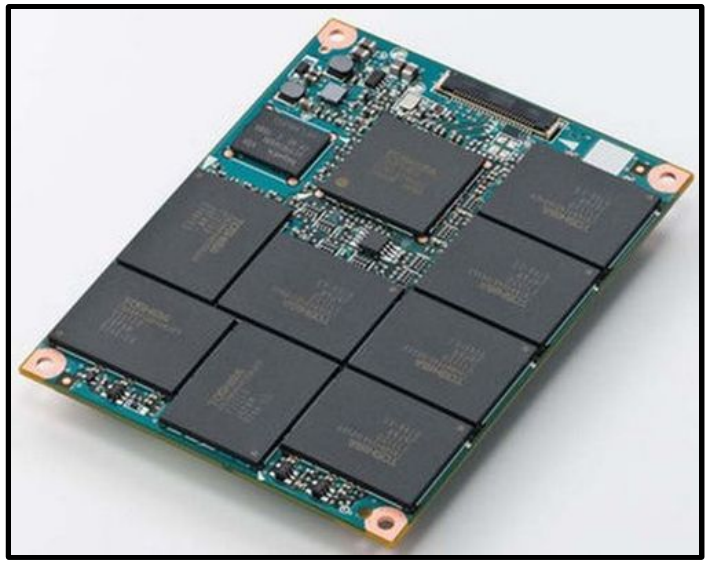
- Para um disco de 7200rpm:
  - **~10ms de seek time**
  - 4.17ms de latência rotacional média
  - Tempo de transferência desprezível

- **Acessos contíguos** têm melhor desempenho
  - Menos movimento da cabeça (seek)

Arquivo e Armazenamento

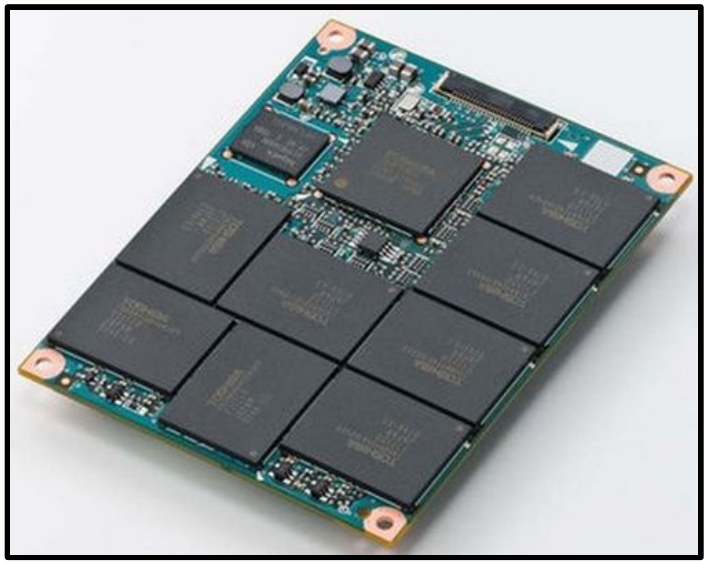
- **Solid State Drive (SSD)**

- Armazenamento em chips de memória não-volátil
- **Não possui partes móveis**
  (como as cabeças dos HDDs)
- Mesma interface que os HDDs, podendo substituir



Arquivo e Armazenamento

- Consome menos energia (se baseado em flash)
- Mais resistente (quedas, movimentos bruscos, etc)
- Ponto negativo número máximo de escritas (se baseado em DRAM)
- **Acesso randômico semelhante ao acesso sequencial**



Arquivo e Armazenamento

	SSD	HDD
Tempo para acesso randômico	~0.1ms	~5-10ms
Custo por capacidade	~US$1.2 – US$2 por GB	~US$0.05 – US$0.10 por GB

Comparação de 2008:

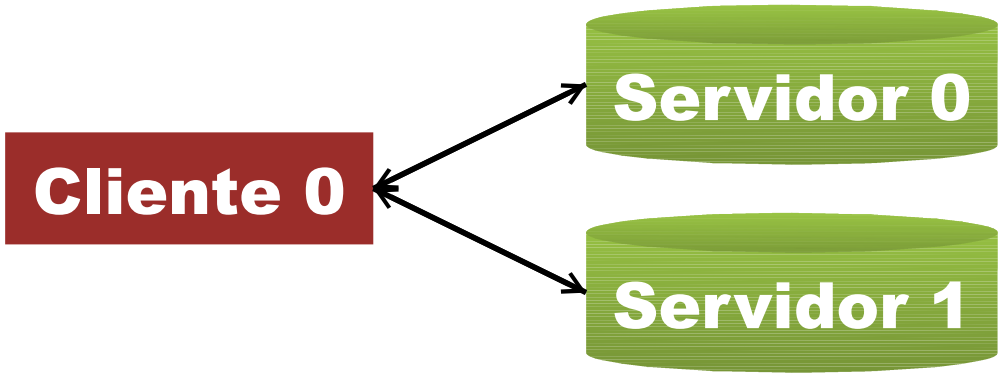
Tipo	Modelo	Capacidade	Preço	Dólares/GB	Tempo de acesso
SSD	Intel X25-M	80GB	US$730	US$9.13	0.085ms
HDD	Seagate 7200 RPM	750GB	US$110	US$0.15	4.2ms

Roteiro

- Introdução
- Arquivo e Armazenamento
- **Sistemas de Arquivos Paralelos**
- Questões de Desempenho
- Tendências
- Conclusão

Sistemas de Arquivos Paralelos

- Permitir acesso a **arquivos remotos**
- Acesso a dados **em paralelo**
  - Foco em HPC



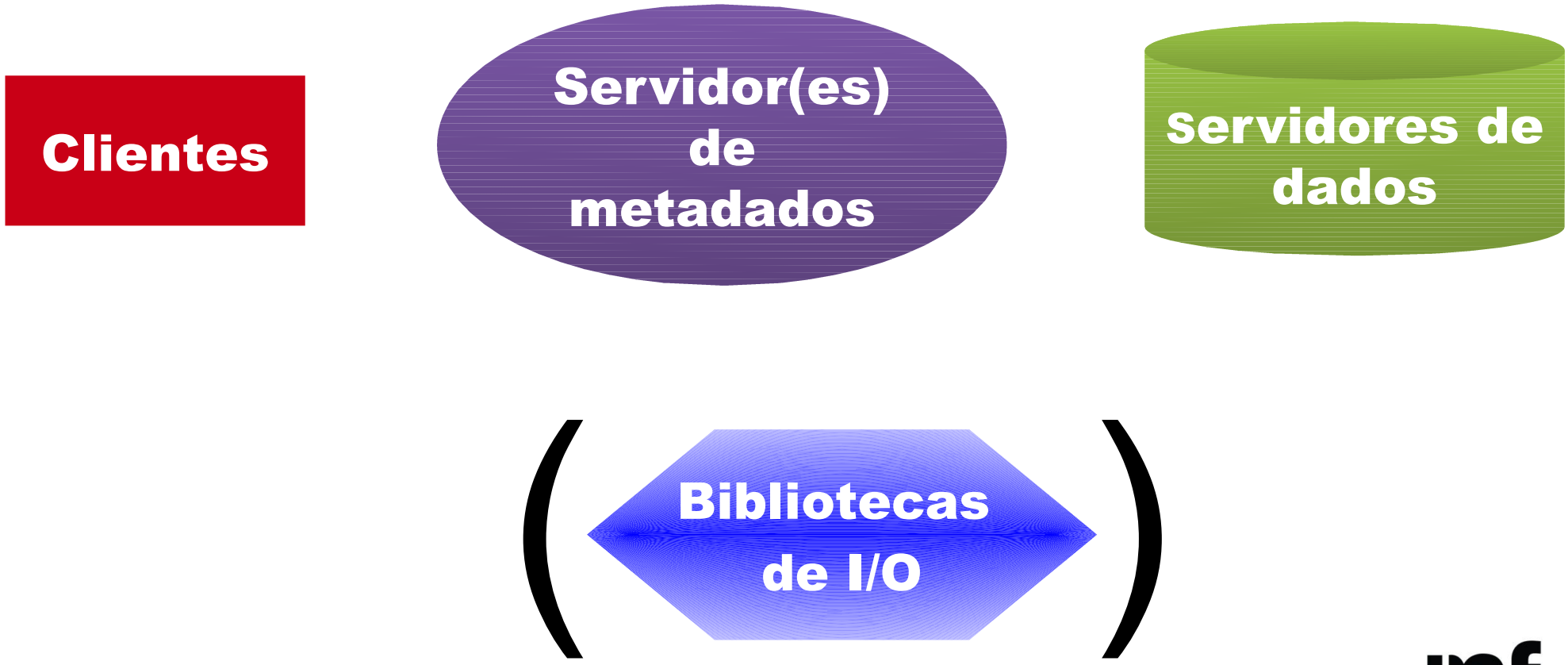
Sistemas de Arquivos Paralelos

- Participação no TOP500

Posição	Máquina	País	Sistema de Arquivos
1	K Computer	Japão	**Lustre**
2	Tianhe-1A	China	Lustre
3	Jaguar	EUA	Lustre
4	Nebulae	China	Solução Mista
5	Tsubame 2.0	Japão	Lustre
6	Cielo	EUA	**Panasas**
15	Intrepid	EUA	**PVFS**

Sistemas de Arquivos Paralelos

- Arquitetura clássica:

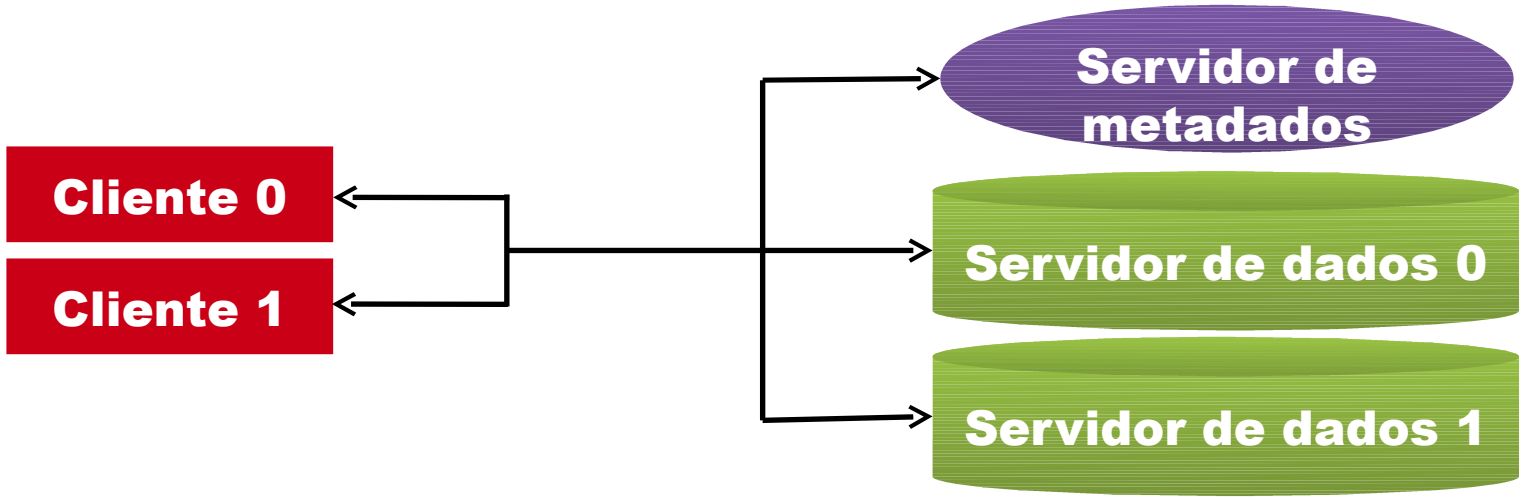


Sistemas de Arquivos Paralelos

- Arquitetura clássica:

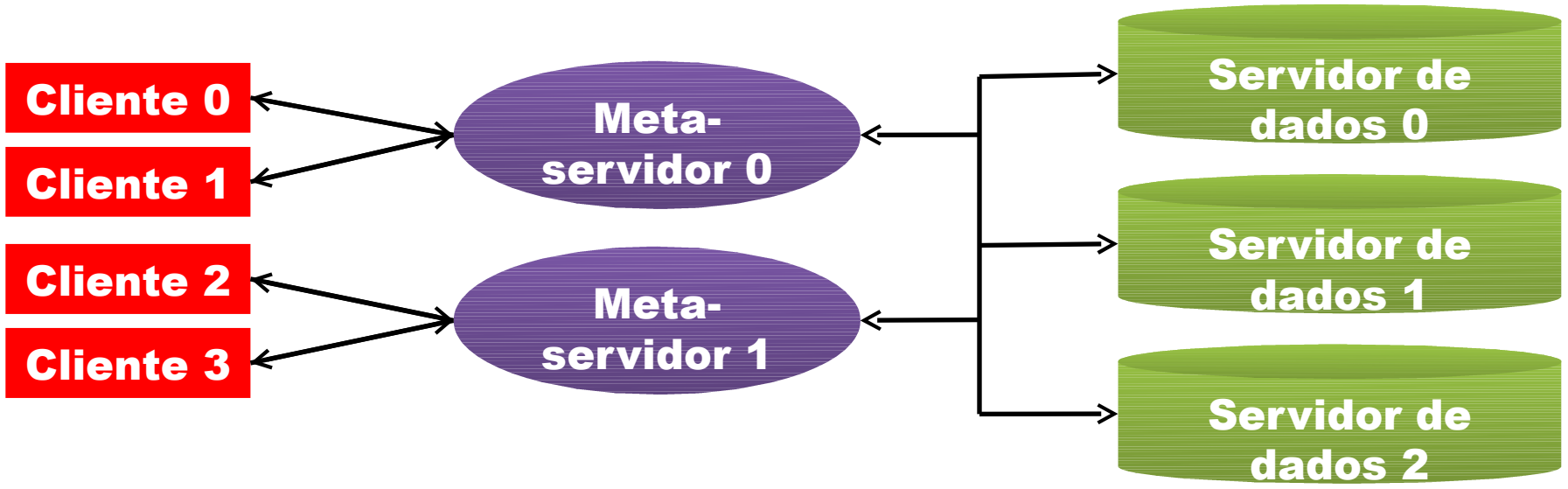
Servidor de Metadados

- Servidores dedicados a metadados
- **Metadados**: informações sobre dados
  - Tamanho, permissões, **localização nos servidores**
- **Clientes consultam** antes de acessar os dados



Servidor de Metadados

- Alguns sistemas **distribuem os metadados**
  - Exemplo: PVFS Contra-exemplo: Lustre
- Distribuir a carga, maior complexidade no sistema

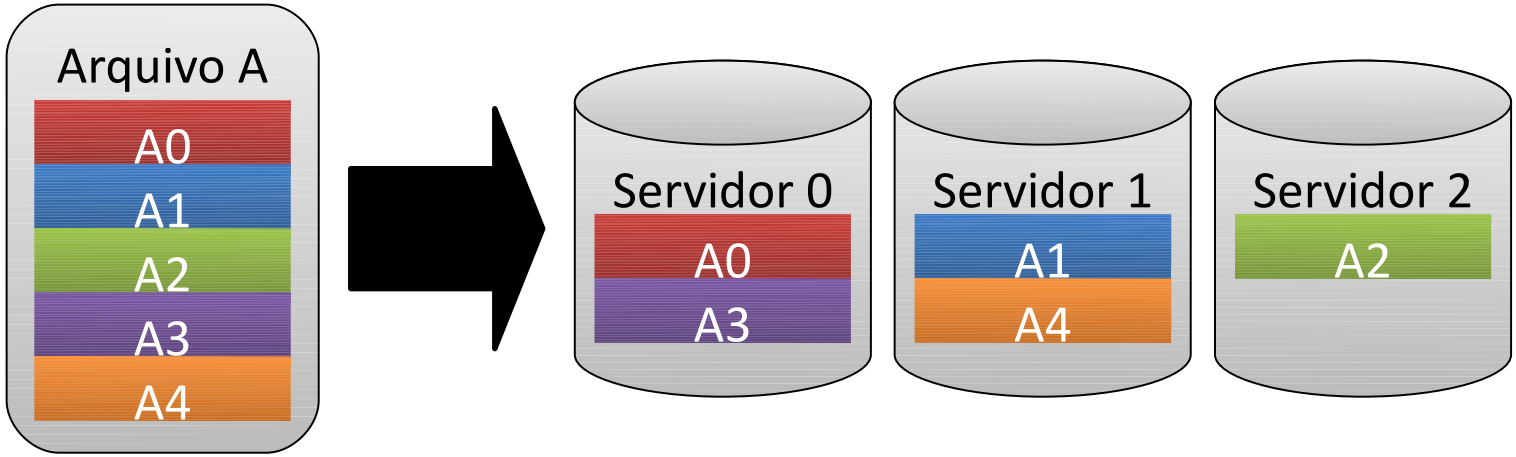


Sistemas de Arquivos Paralelos

- Arquitetura clássica:

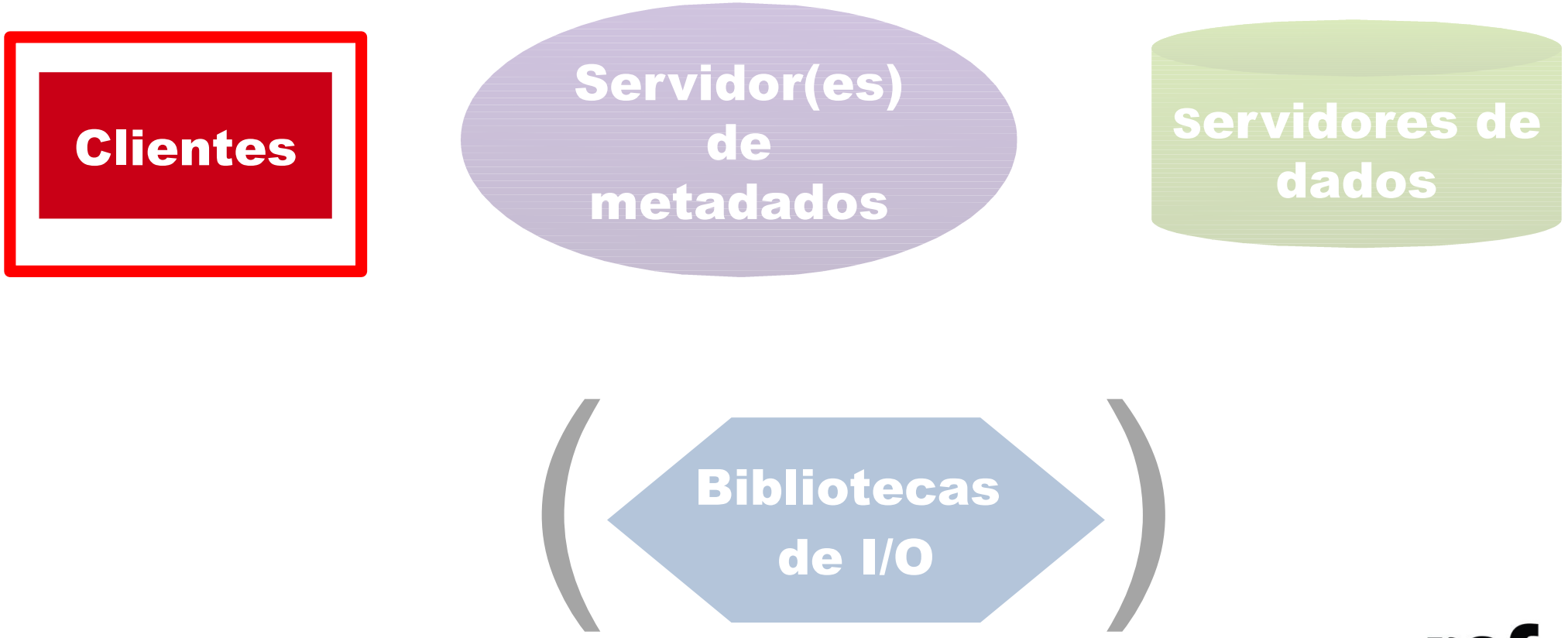
Servidores de Dados

- **Dados são distribuídos entre servidores**
  - Arquivo é fatiado (**striping**) e as fatias são distribuídas
- Tamanho das fatias (stripe size) e número de servidores **afetam o desempenho**



Sistemas de Arquivos Paralelos

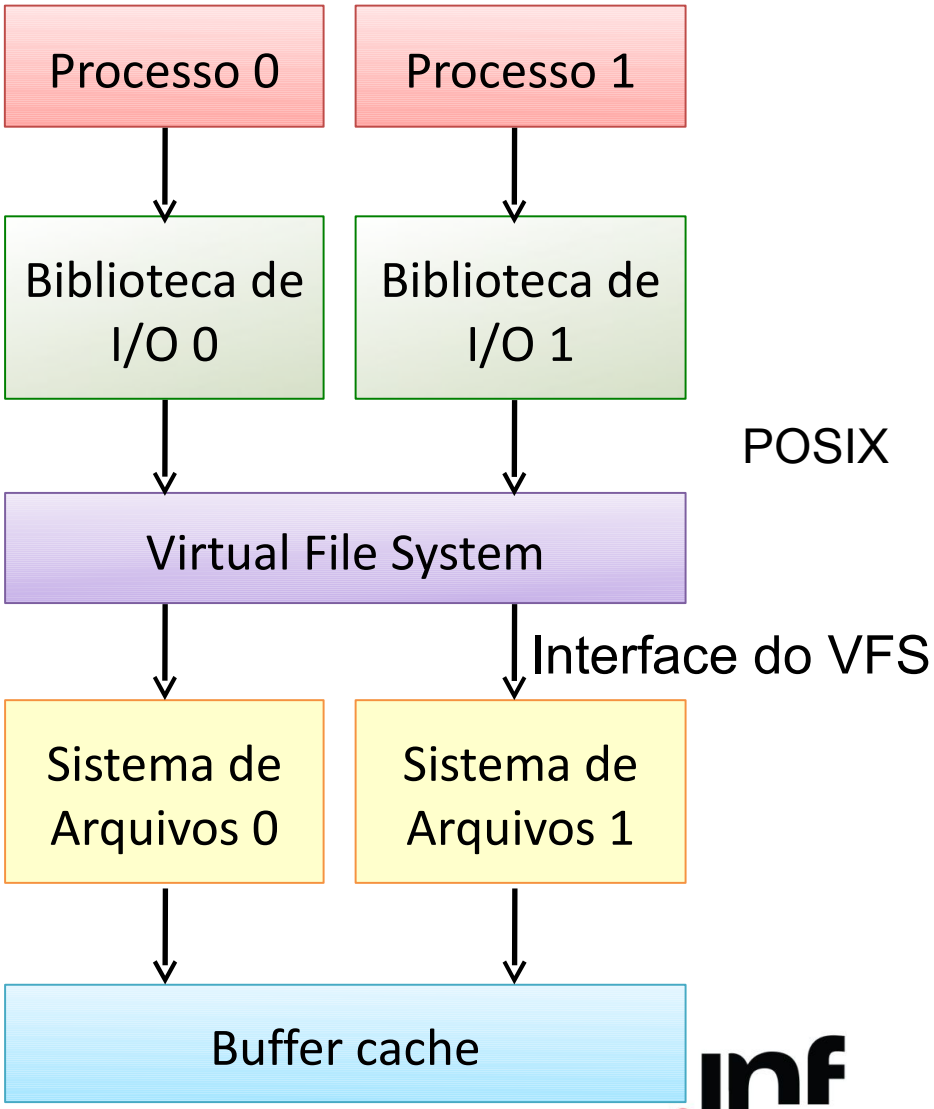
- Arquitetura clássica:



Clientes

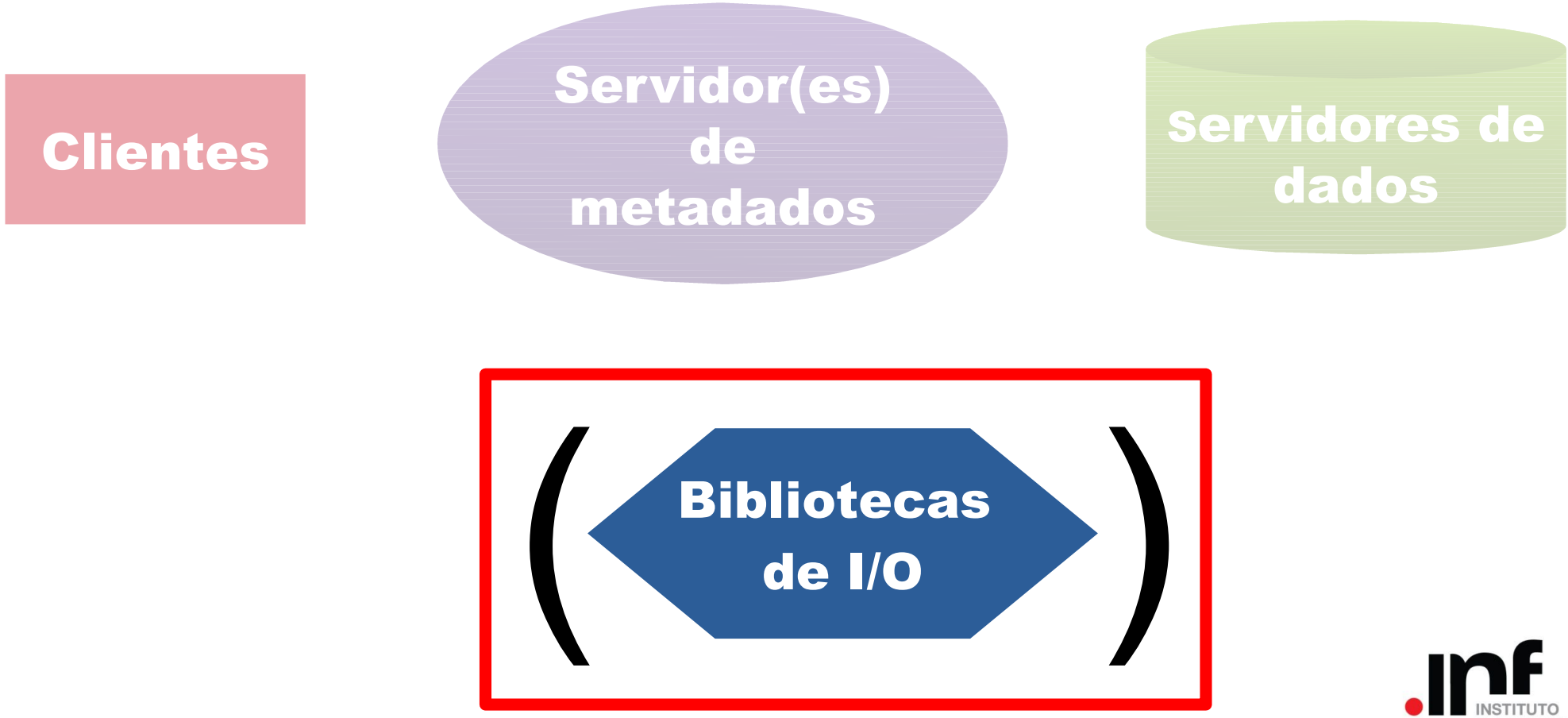
- Camadas no **cliente**:

- Sistema de Arquivos deve ser **transparente**
  - Aplicação não sabe se arquivo é remoto



Sistemas de Arquivos Paralelos

- Arquitetura clássica:



Bibliotecas de I/O

- SAP é mais otimizado para **algumas situações**
- **Desempenho** depende de como a aplicação acessa
- **Bibliotecas podem fazer otimizações** nos acessos

Roteiro

- Introdução
- Arquivo e Armazenamento
- Sistemas de Arquivos Paralelos
- **Questões de Desempenho**
- Tendências
- Conclusão

Questões de Desempenho

- Acessos Pequenos e Esparsos
- Arquivos Pequenos e/ou Numerosos
- Acessos Fora de Ordem
- Cache de Dados no Cliente

Questões de Desempenho

- **Acessos Pequenos e Esparsos**
- Arquivos Pequenos e/ou Numerosos
- Acessos Fora de Ordem
- Cache de Dados no Cliente

Acessos Pequenos e Esparsos

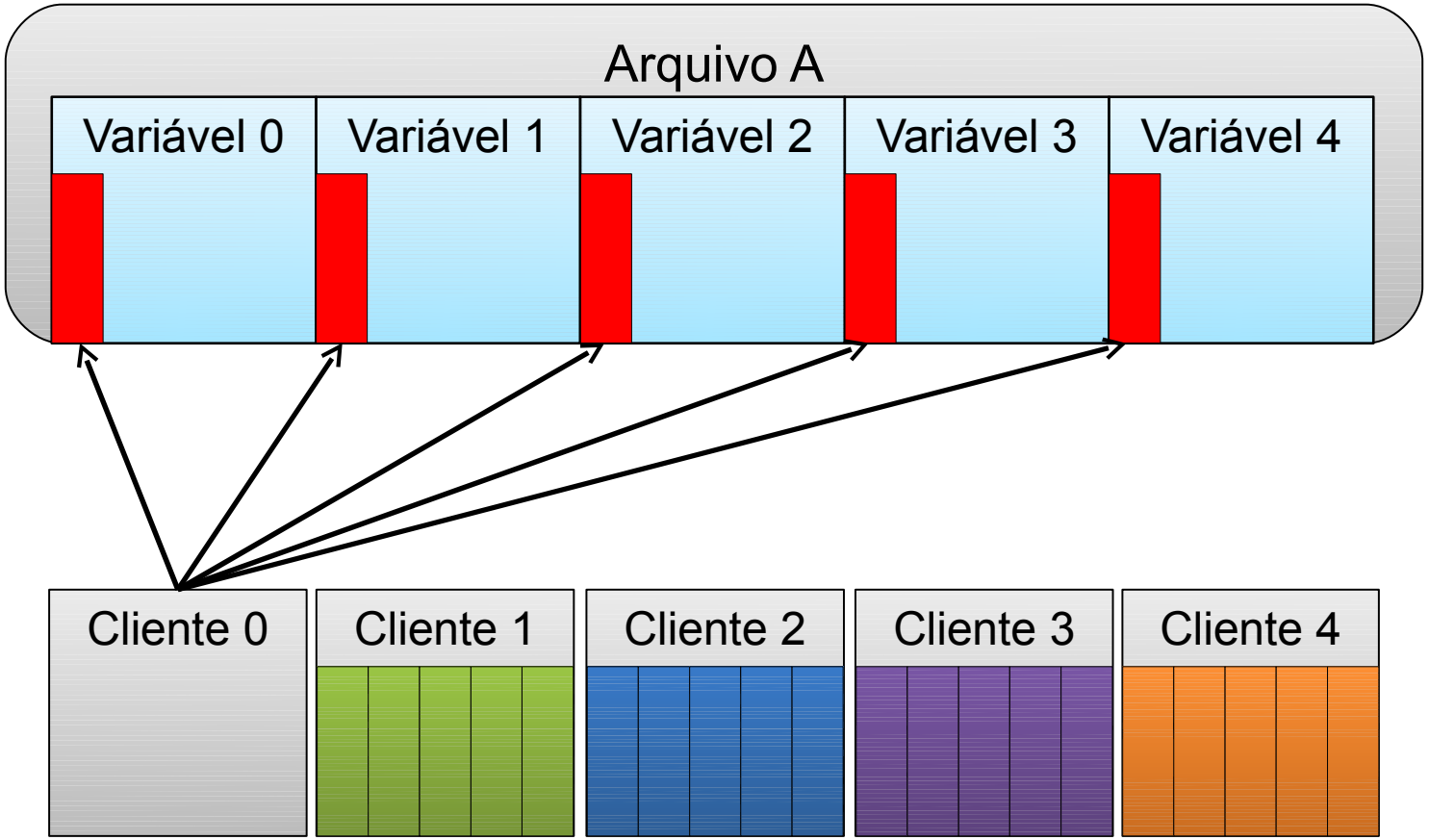
- Padrão de acesso **comum em aplicações científicas**
- Pior desempenho
  - Dados os custos fixos de um acesso

Acessos Pequenos e Esparsos

- Exemplo: **checkpoint**
  - Porção do arquivo = valor de uma variável nos processos

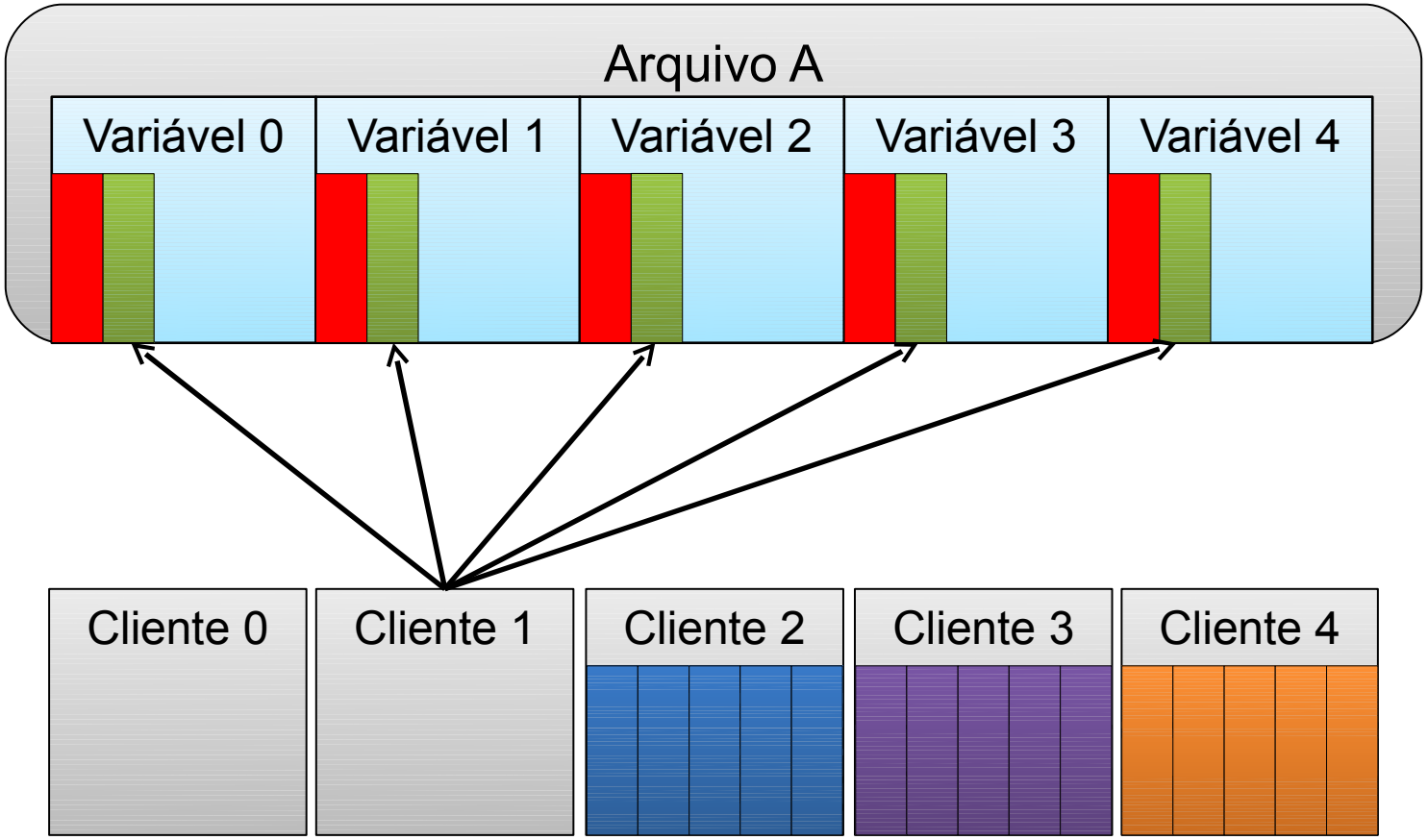
Acessos Pequenos e Esparsos

- Exemplo: **checkpoint**
  - Porção do arquivo = valor de uma variável nos processos



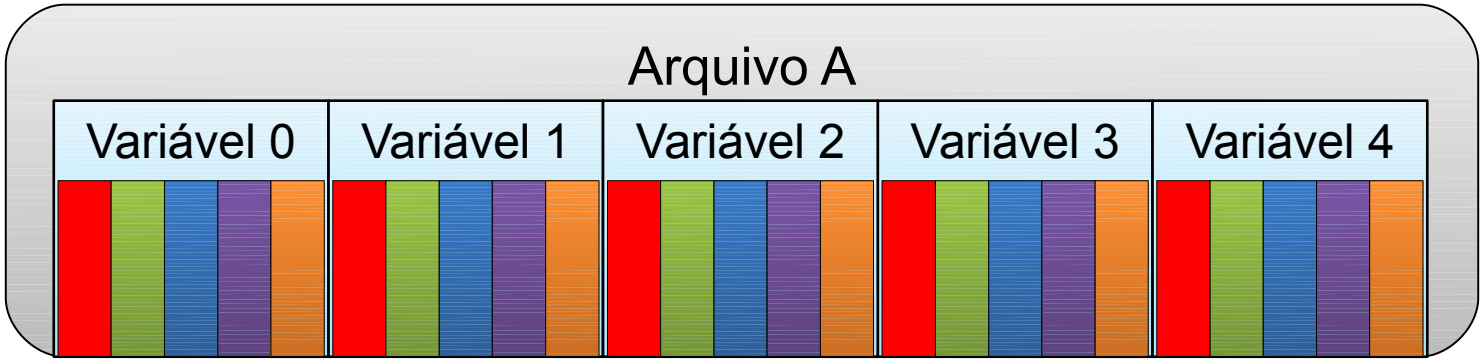
Acessos Pequenos e Esparsos

- Exemplo: **checkpoint**
  - Porção do arquivo = valor de uma variável nos processos



Acessos Pequenos e Esparsos

- Exemplo: **checkpoint**
  - Porção do arquivo = valor de uma variável nos processos

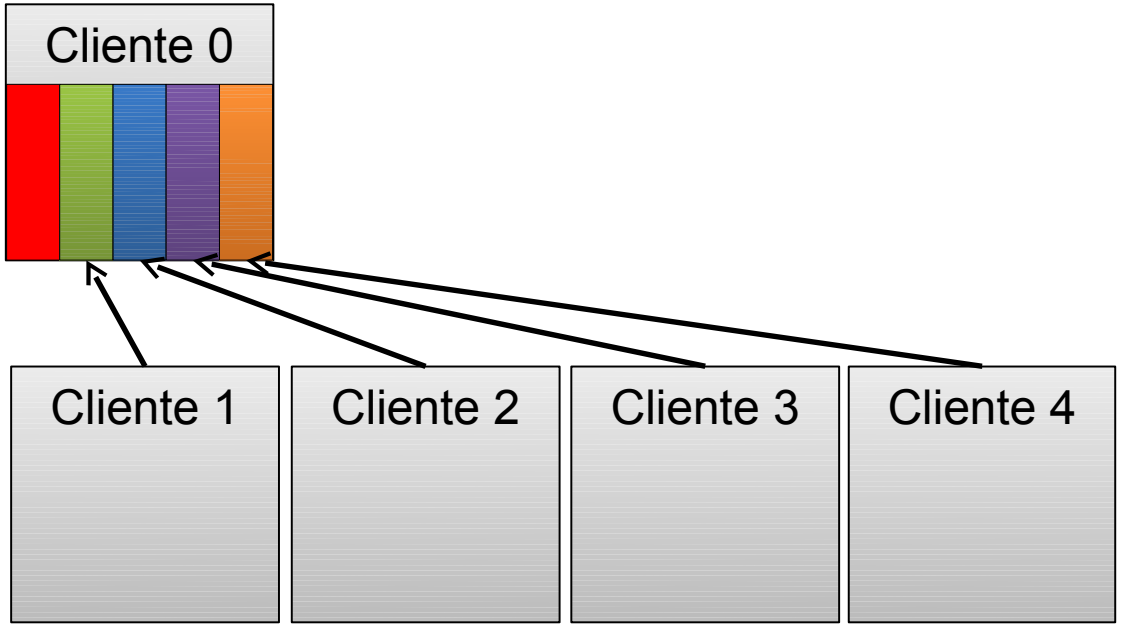


Acessos Pequenos e Esparsos

- Solução comum: **operações coletivas**
  - Biblioteca reorganiza acessos (com sincronização)
  - Exemplo: MPI-IO (estratégia em duas fases)

Acessos Pequenos e Esparsos

Fase 1

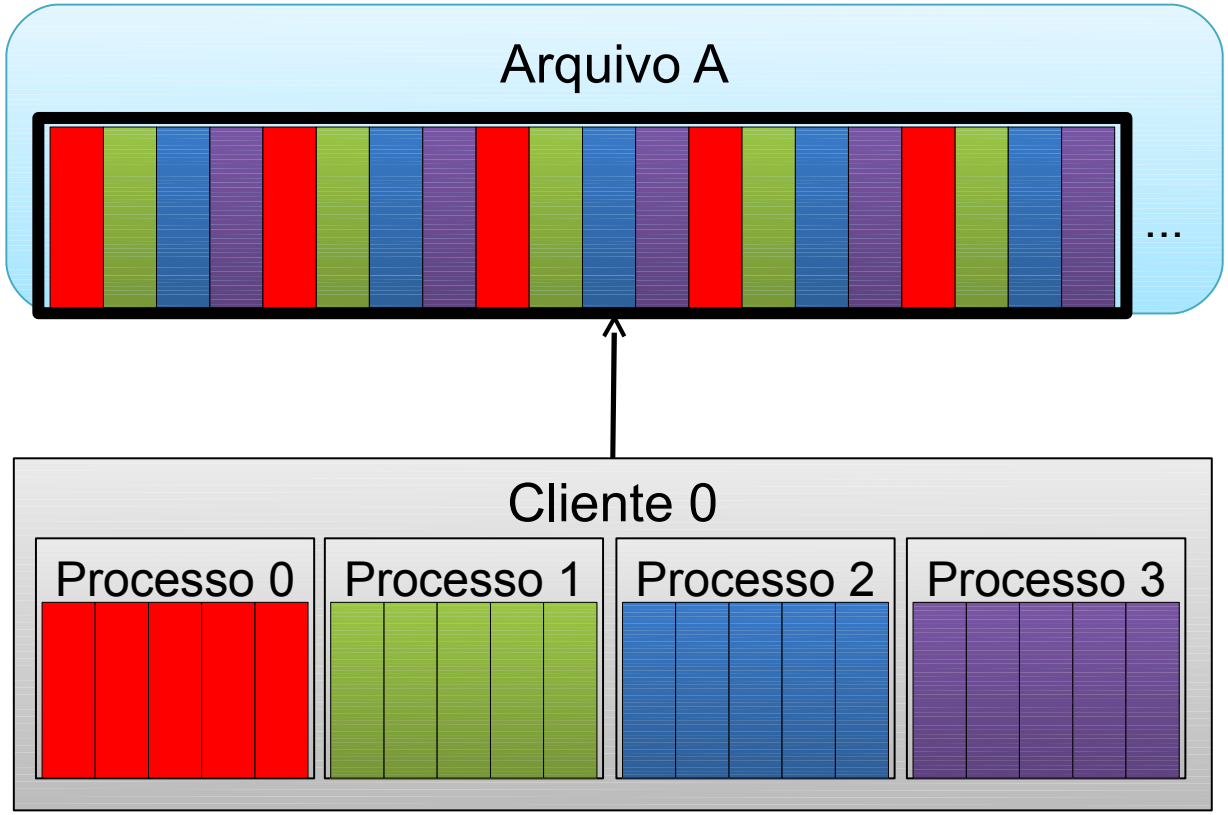


Acessos Pequenos e Esparsos

Fase 2

Acessos Pequenos e Esparsos

- Mesma idéia pode ser empregada **intrá-no**
  - Processos concorrentes



Questões de Desempenho

- Acessos Pequenos e Esparsos
- **Arquivos Pequenos e/ou Numerosos**
- Acessos Fora de Ordem
- Cache de Dados no Cliente

Arquivos Pequenos e/ou Numerosos

- Situações com **arquivos pequenos**
  - Desempenho ruim
  - Evidencia sobrecusto da criação (operação no meta-servidor)
- Situações com **muitos arquivos**
  - Desempenho ruim
  - Sobrecarga nos meta-servidores

Arquivos Pequenos e/ou Numerosos

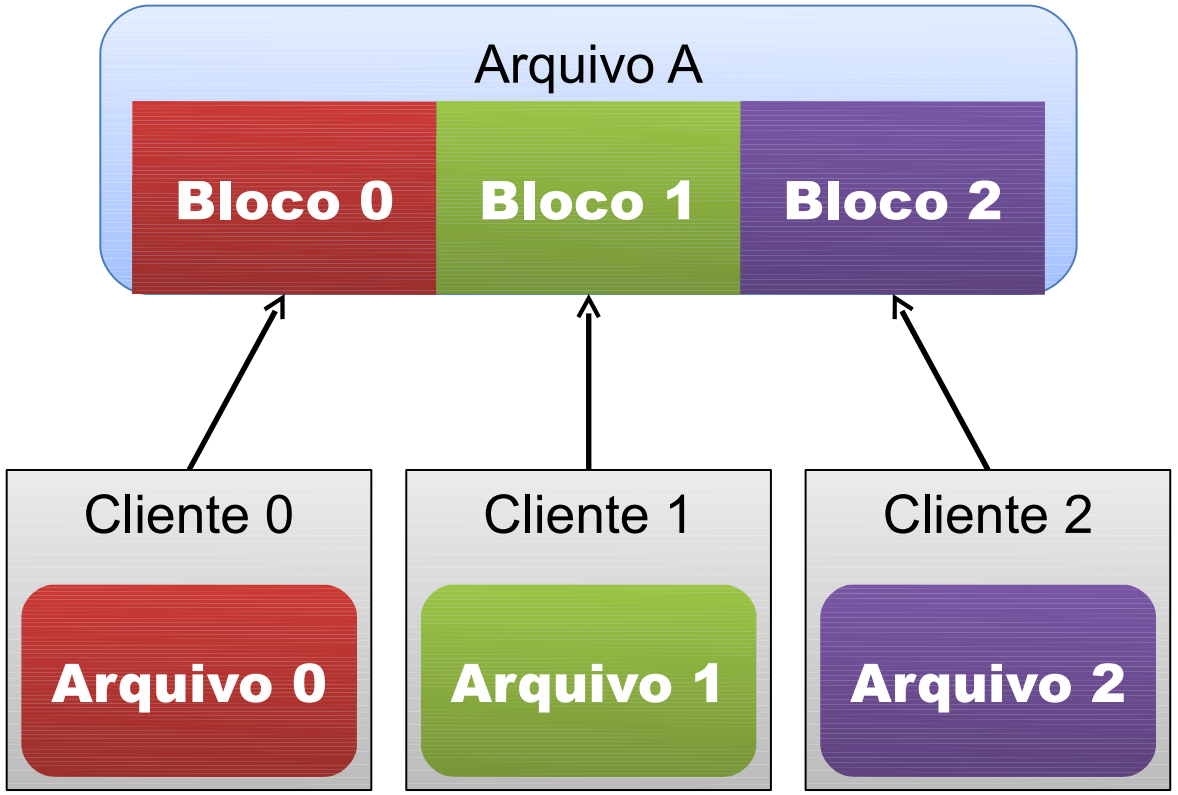
- Situações com arquivos pequenos
- Situações com muitos arquivos

Desempenho depende dos metadados!

- Melhorar desempenho: otimizar operações sobre metadados

Arquivos Pequenos e/ou Numerosos

- Alternativa: biblioteca para mapeamento
  - **muitos arquivos virtuais -> poucos arquivos reais**

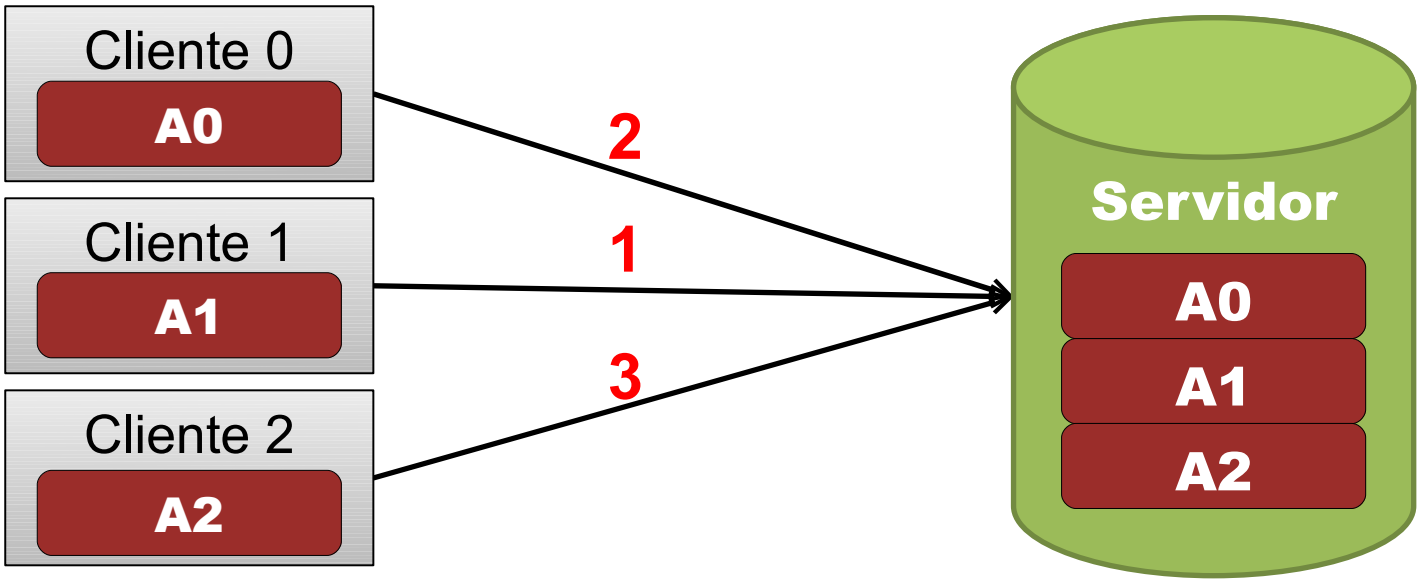


Questões de Desempenho

- Acessos Pequenos e Esparsos
- Arquivos Pequenos e/ou Numerosos
- **Acessos Fora de Ordem**
- Cache de Dados no Cliente

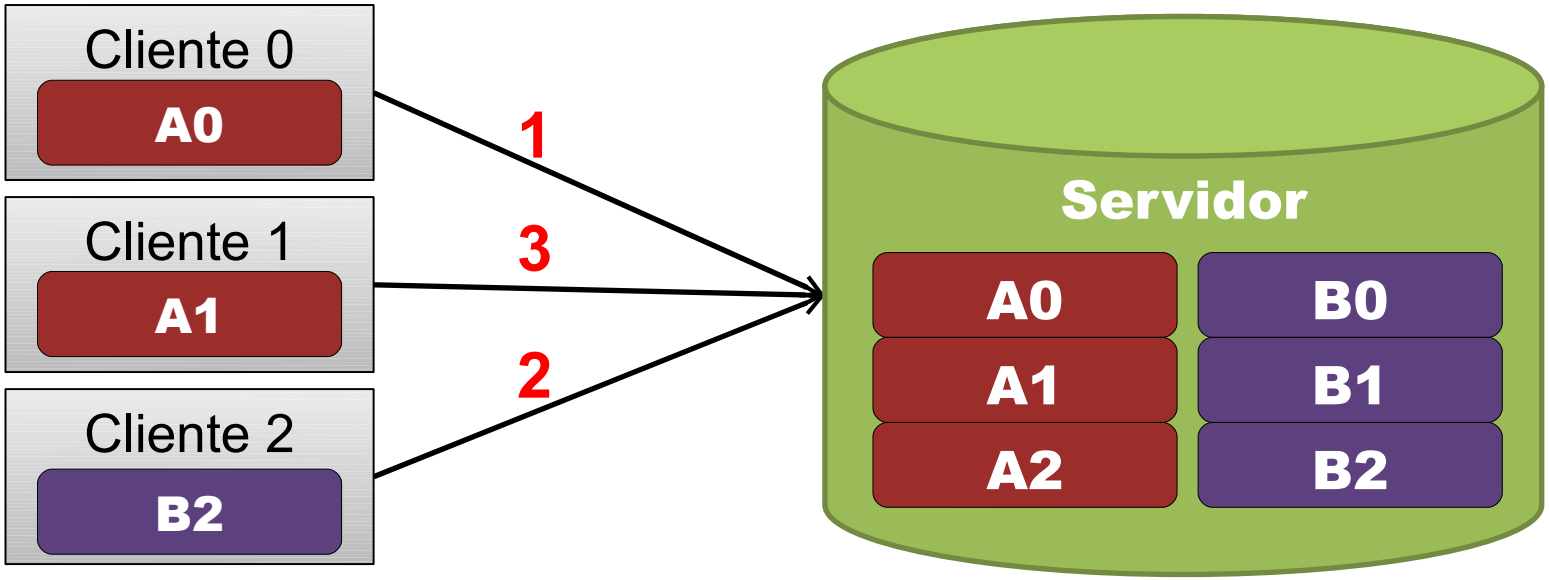
Acessos Fora de Ordem

- **Acessos contíguos** têm melhor desempenho
- Então a ordem das requisições importa
- Biblioteca pode **reorganizar/agrupar acessos**



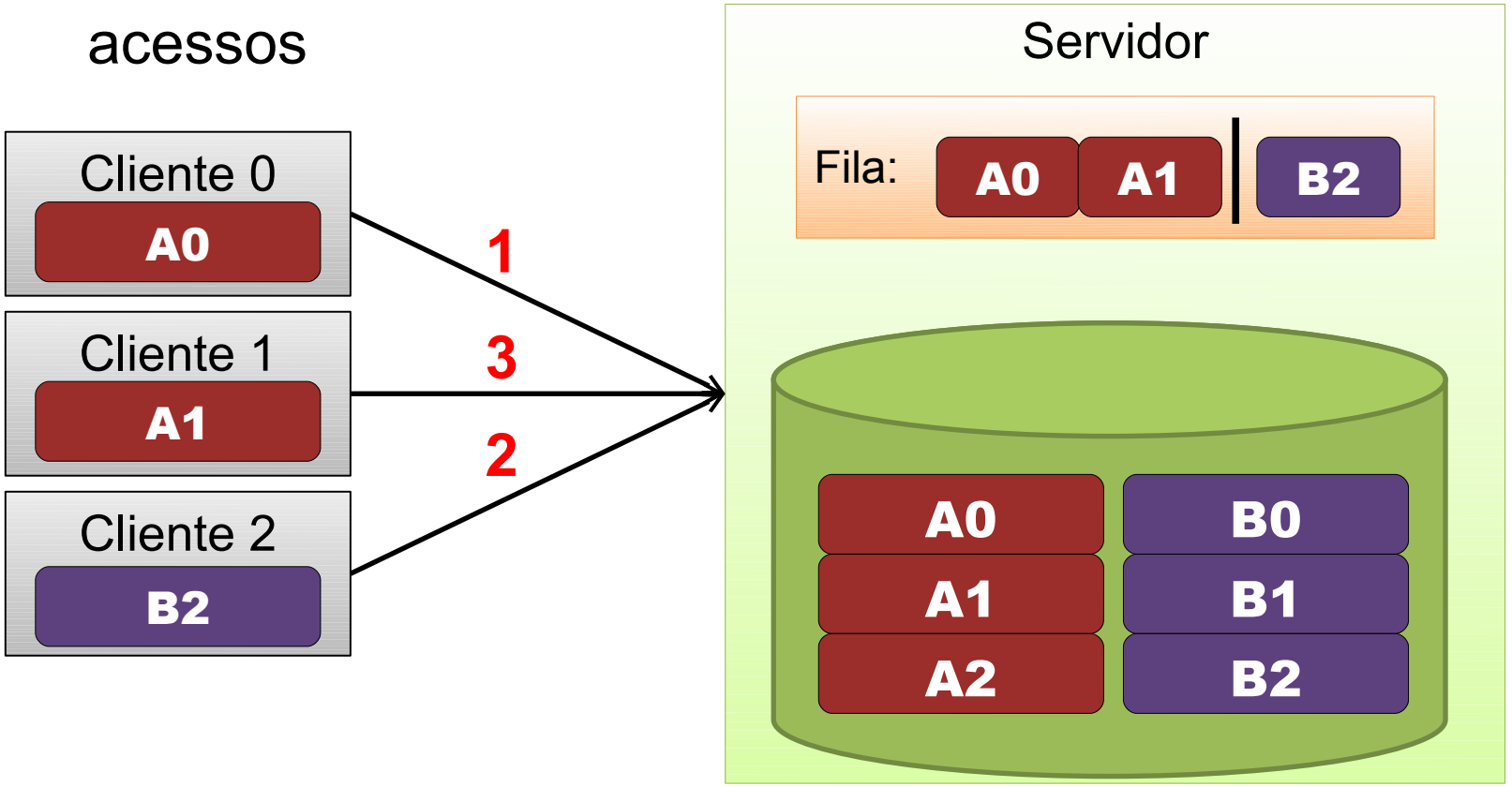
Acessos Fora de Ordem

- É comum que **aplicações compartilhem o SAP**
- Otimizações individuais podem ser prejudicadas



Acessos Fora de Ordem

- **Escalonamento** de operações de entrada e saída
  - Garantir justiça e tempo de resposta, reorganizar e agrupar acessos

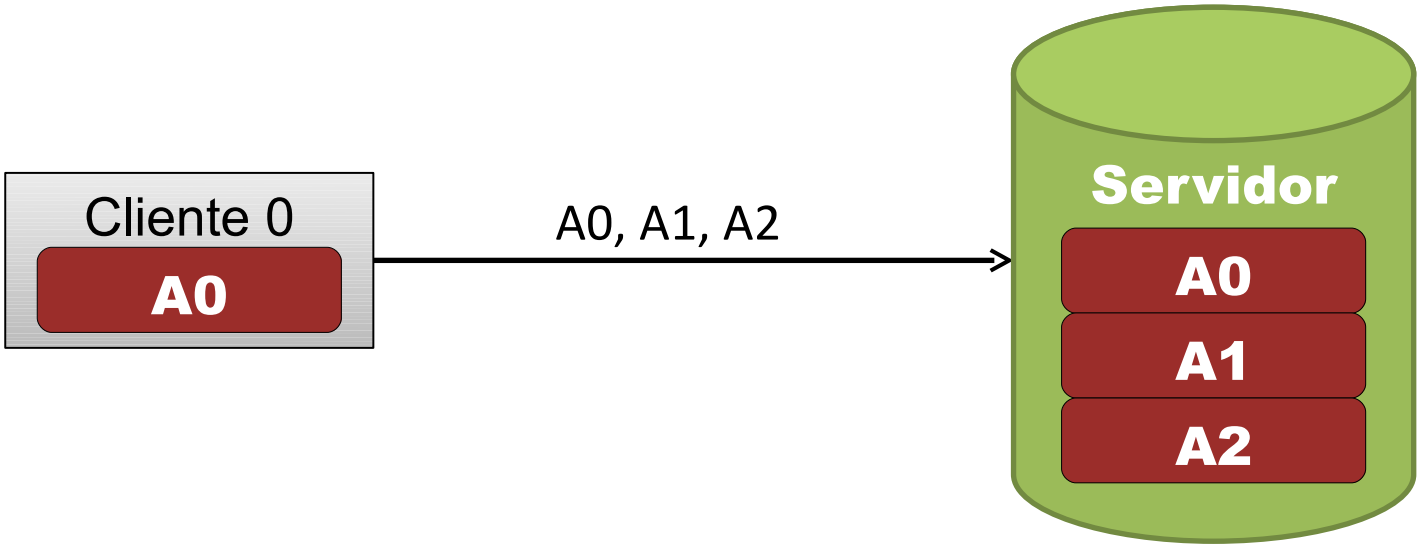


Questões de Desempenho

- Acessos Pequenos e Esparsos
- Arquivos Pequenos e/ou Numerosos
- Acessos Fora de Ordem
- **Cache de Dados no Cliente**

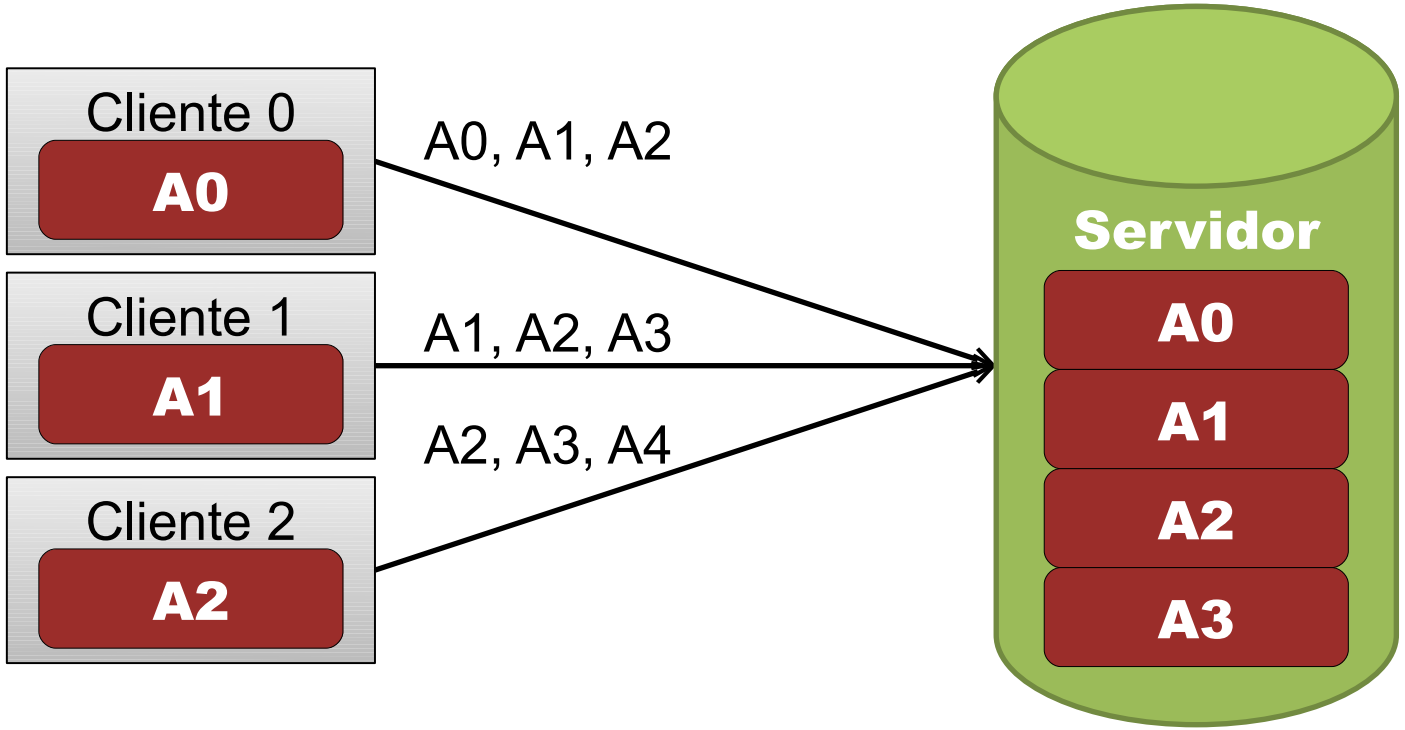
Cache de Dados no Cliente

- Usada para **esconder latência** do SAP
- Para melhorar desempenho, **prefetching**
  - Buscar antes o que será acessado a seguir



Cache de Dados no Cliente

- Prefetching é mais útil quando inteligente
- **Incluir conhecimento da aplicação**



Cache de Dados no Cliente

- **Application-aware Prefetching**
- Extração do padrão de acesso da aplicação
  - Pré-execução
  - Com *hints* do desenvolvedor
  - ...

Questões de Desempenho

- **Diversos outros problemas/soluções**
  - Alinhamento das requisições ao tamanho do stripe
  - Otimizações específicas para checkpoint
  - Minimizar número de conexões entre cliente e servidores
  - Escolher dinamicamente biblioteca de I/O
  - Reconfiguração automática do sistema de arquivos
  - ...

Roteiro

- Introdução
- Arquivo e Armazenamento
- Sistemas de Arquivos Paralelos
- Questões de Desempenho
- **Tendências**
- Conclusão

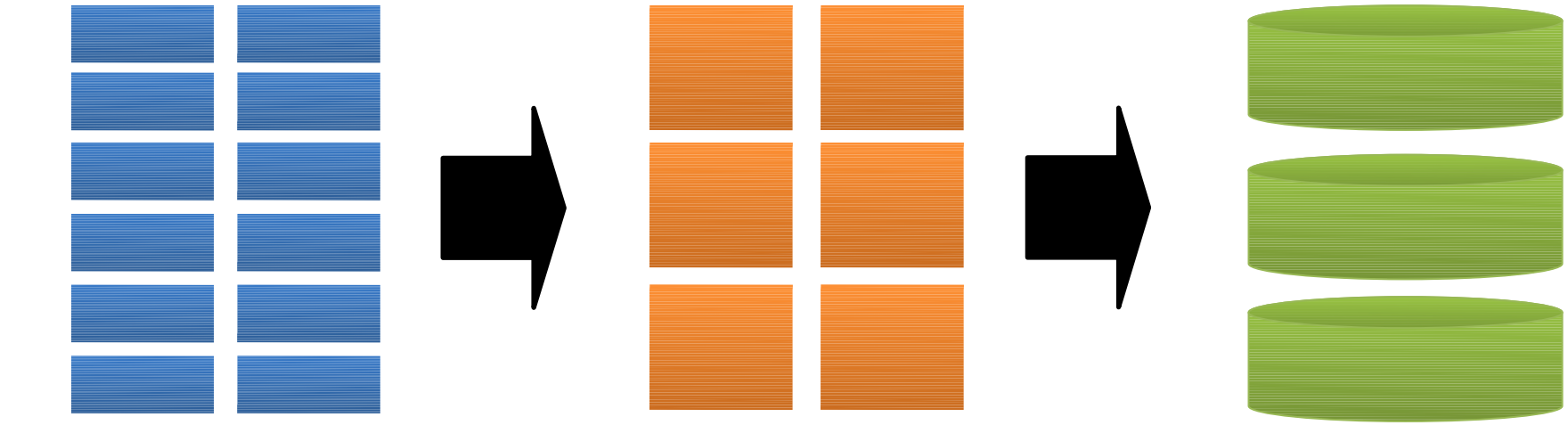
Tendências

- Sistemas de Arquivos e Bibliotecas deverão usar **informações sobre a aplicação**

- Maior integração entre as camadas
- Suportar **novas tecnologias**
- Abandonar semânticas sequenciais

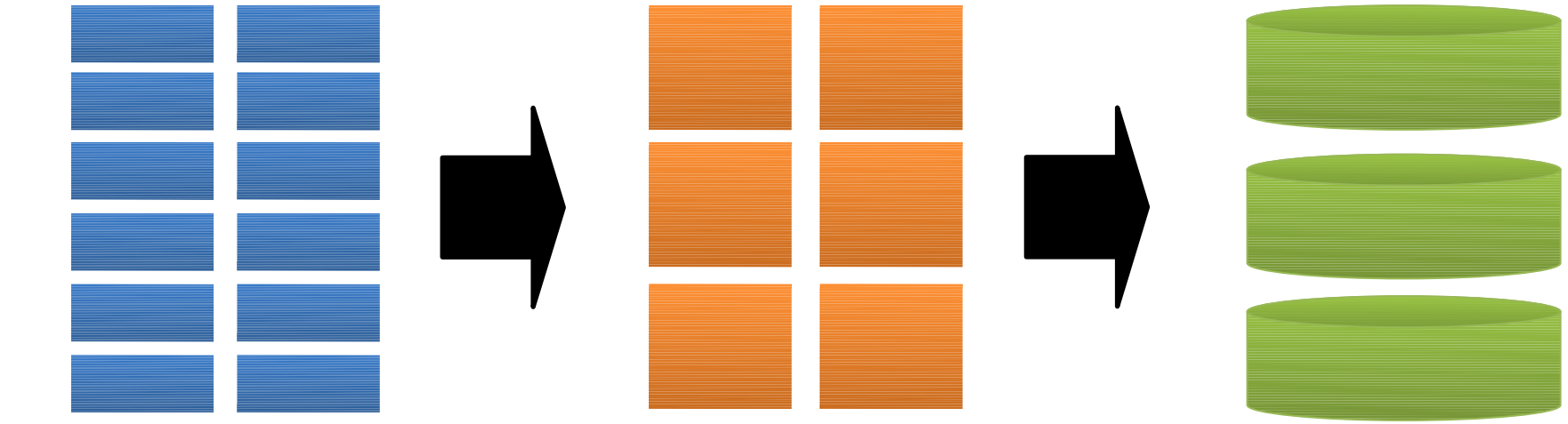
Encaminhamento de I/O

- **Diminuir a concorrência** no SAP
  - Sistemas de **enorme escala**
- Simplifica nós de processamento



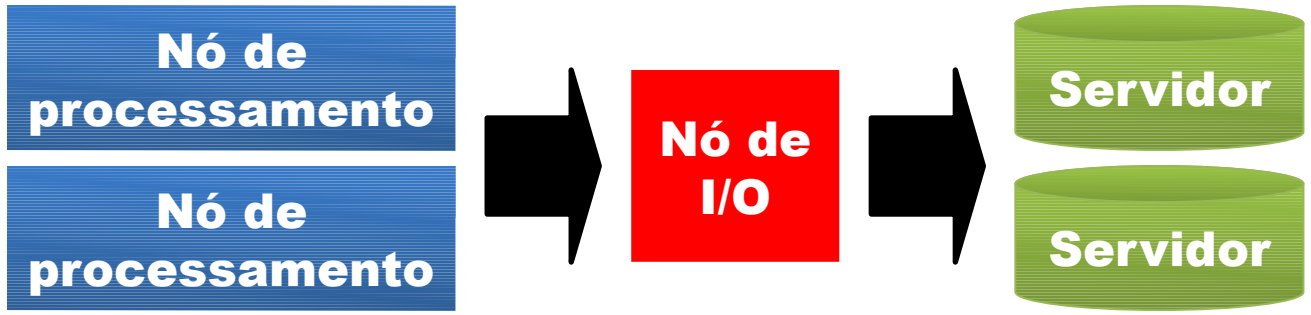
Encaminhamento de I/O

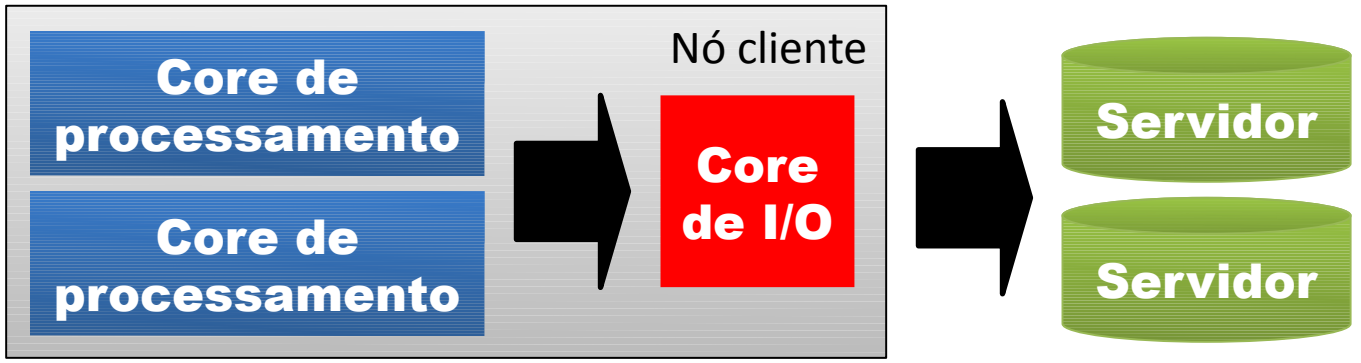
- Maior contribuição: **local para otimizações**
  - Agrupamento e reordenamento de requisições, escalonamento, prefetching, ...



Encaminhamento de I/O

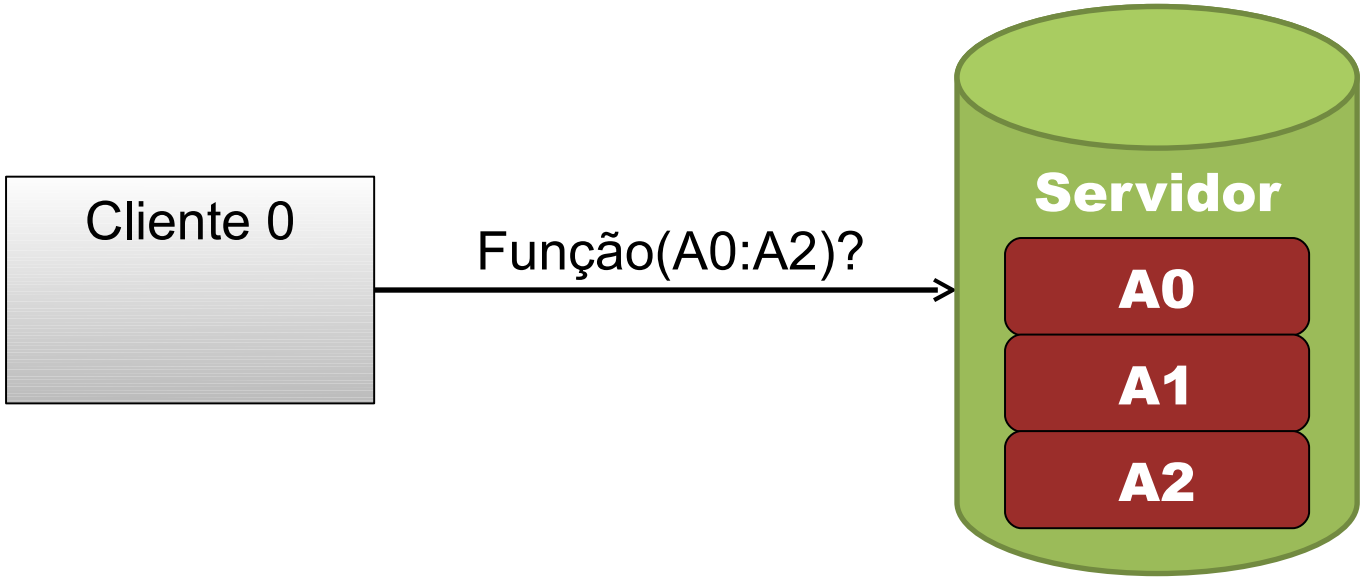
- Mesma idéia pode ser empregada **intra-nó**:





Armazenamento Ativo

- **Poder de processamento** subutilizado em servidores
- Capacitá-los a efetuar **operações sobre os dados**



Roteiro

- Introdução
- Arquivo e Armazenamento
- Sistemas de Arquivos Paralelos
- Questões de Desempenho
- Tendências
- **Conclusão**

Conclusão

- **I/O é mais lento que processamento**
  - Desempenho das aplicações limitado por essas operações
- Sistemas de arquivos paralelos
  - Compartilhamento entre instâncias de aplicações em cluster
  - **Esconder latência do I/O através de acesso paralelo**

Conclusão

- **I/O é um problema crítico** no caminho ao exascale
- Desempenho depende de muitos fatores
  - Diversas otimizações para situações específicas
- No futuro: **mais inteligência no sistema de arquivos**
  - Uso de informações sobre a aplicação

Desafios de E/S em Ambientes de Grande Escala

Philippe O. A. Navaux
Francieli Zanon Boito
Rodrigo Virote Kassick

Grupo de Processamento Paralelo e Distribuído (GPPD)
Universidade Federal do Rio Grande do Sul (UFRGS)

ERAD-SP 2011 – São José dos Campos/SP

Referências

- [Ávila 2005] R. Ávila. "Uma Proposta de Distribuição do Servidor de Arquivos em Cluster" (tese de doutorado na UFRGS), 2005
- [Boito et al. 2011] F. Boito et al. "I/O Performance of a Large Atmospheric Model using PVFS" (artigo na conferência RenPar), 2011
- [Brandt et al. 2003] A. Brandt, E. Miller, D. Long e L. Xue. "Efficient Metadata Management in Large Distributed Storage Systems" (artigo na Conferência IEEE NASA MSS), 2003
- [Buyya 1999] Rajkumar Buyya (editor). "High Performance Cluster Computing: Programming and Applications" (livro), 1999
- [Byna et al. 2008] S. Byna, Y. Chen, X. Sun, R. Thakur e W. Gropp. "Parallel I/O prefetching using MPI file caching and I/O signatures"(artigo na Conferência SC), 2008
- [Carns et al. 2009] P. Carns, S. Lang, R. Ross, M. Vilayannur, J. Kunkel e T. Ludwig "Small-file access in parallel file systems" (artigo na Conferência IEEE IPDPS), 2009
- [Coulouris et al. 2007] G. Coulouris et al. "Sistemas Distribuídos: Conceitos e Projeto" (livro), 2007

Referências

- [Dias et al. 2010] P. Dias et al. “Análise de Desempenho da Versão Paralela do OLAM” (relatório do LNCC), 2010
- [Dongarra et al. 2011] J. Dongarra et al. “The International Exascale Software Project Roadmap” (artigo no IJHPCA), 2011. www.exascale.org
- [Frings et al. 2009] W. Frings, F. Wolf e V. Petkov. “Scalable massively parallel I/O to task-local files” (artigo na Conferência SC), 2009.
- [Latham et al. 2004] R. Latham, N. Miller, R. Ross e P. Carns. “A Next-Generation Parallel File System for Linux Clusters” (artigo na publicação LinuxWorld), 2004
- [Lebre et al. 2006] A. Lebre, Y. Denneulin, G. Huard e P. Sowa. “I/O scheduling service for multi-application clusters” (artigo na Conferência IEEE Cluster), 2006
- [Liao et al. 2007] W. Liao, A. Ching, K. Coloma, A. Choudhary e M. Kandemir. “Improving MPI independent write-performance using a two-stage write-behind buffering method” (artigo na Conferência IEEE IPDPS), 2007

Referências

- [Ohta et al. 2009] K. Ohta, H. Matsuba e Y. Ishikawa. "Improving parallel write by node-level request scheduling" (artigo na Conferência ACM/IEEE CCGrid), 2009
- [Oliveira, 2011] J. Palazzo M. de Oliveira, http://palazzo.pro.br/disco.htm (acessado em julho de 2011)
- [Patterson , Hennessy 2009] D. Patterson e J. Hennessy. "Computer Organization and Design: The Hardware / Software Interface" (livro), 2009
- [Piernas et al. 2007] J. Piernas, J. Nieplocha e E. Felix. "Evaluation of active storage strategies for the Lustre parallel file system" (artigo na Conferência SC), 2007
- [Polte et al. 2008] M. Polte, J. Simsa, G. Gibson. "Comparing performance of solid state devices and mechanical disks" (artigo no Petascale Data Storage Workshop), 2008
- [Qian et al. 2009] Y. Qian, E. Barton, T. Wang, N. Puntambekar e A. Dilger. "A novel network request scheduler for a large-scale storage system" (artigo na publicação Computer Science – Research and Development), 2009
- [Tanenbaum 2008] A. Tanenbaum. "Modern Operating Systems" (livro), 2008

Referências

- [Tanenbaum e Steen 2007] A. Tanenbaum e M. Steen. "Sistemas Distribuídos: Princípios e Paradigmas" (livro), 2007
- [Thakur et al 1998] R. Thakur, W. Gropp e E. Lusk. "Data sieving and Collective I/O in ROMIO" (artigo no Symposium on the Frontiers of Massively Parallel Computation), 1998
- [TOP500] www.top500.org, acessado em 2011
- [Vishwanath et al. 2010] V. Vishwanath, M. Hereld, K. Iskra, D. Kimpe, V. Mozorov, M. Papka, R. Ross e K. Yoshii. "Accelerating I/O forwarding in IBM Blue Gene/P systems" (artigo na Conferência SC), 2010
- [Zhang et al. 2009] X. Zhang, S. Jiang e K. Davis. "Making resonance a common case: A high-performance implementation of collective I/O on parallel file systems"(artigo na Conferência IEEE IPDPS), 2009
- [Zhang et al. 2010] X. Zhang, K. Davis, S. Jiang. "IOrchestrator: improving the performance of multi-node I/O systems via inter-server coordination" (artigo na Conferência SC), 2010