Uma proposta para migração de páginas no Linux

Guilherme A. A. Tesser (HP/PUCRS)
Avelino F. Zorzo (PUCRS)

PUCRS/HP – Porto Alegre - Brazil

Sumário

- Introdução
- Escalonador do Linux
- Balanceamento de carga em máquinas NUMA
- Balanceamento de carga do Linux
- Balanceamento de carga multi-nível
- Migração de páginas
- Resultados
- Conclusão

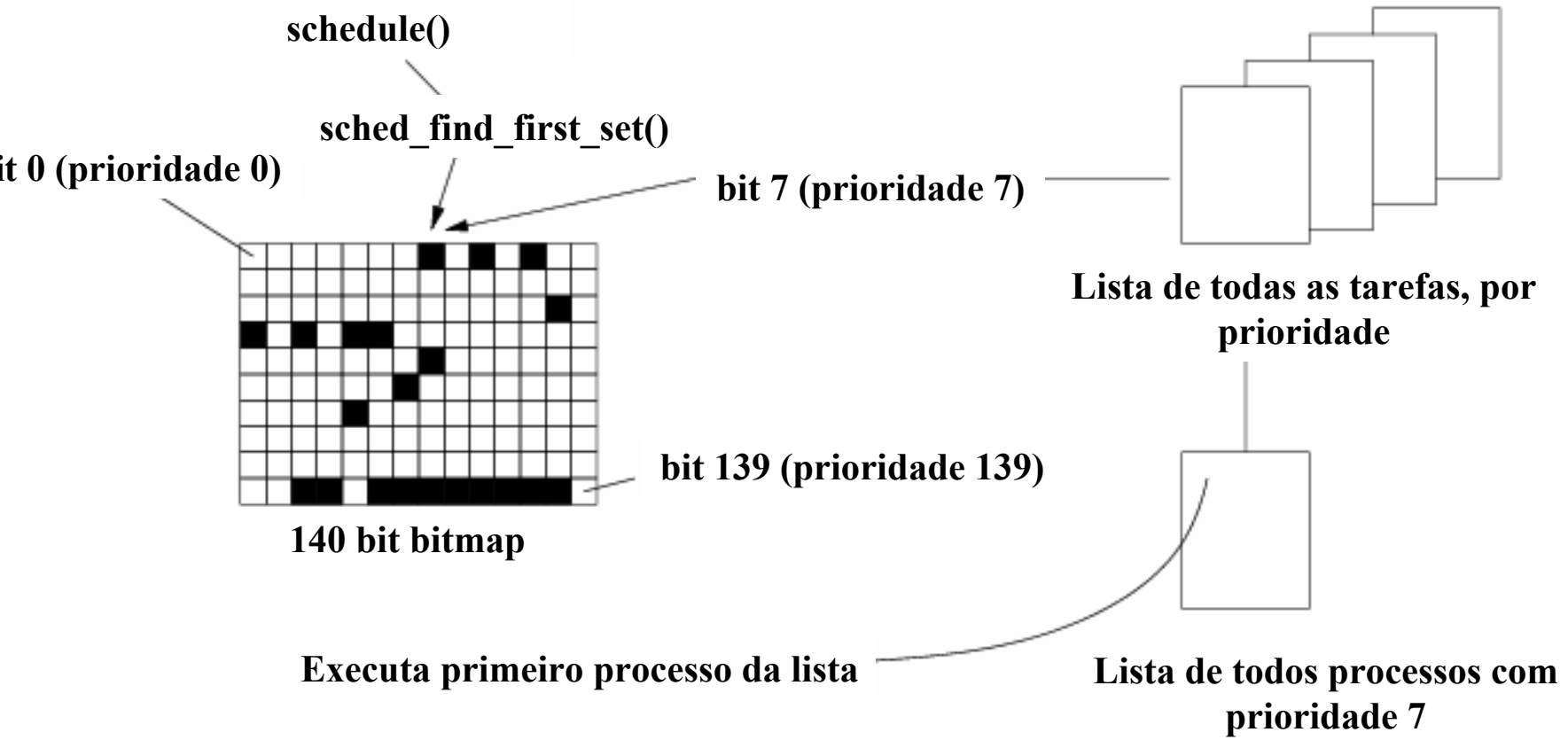
Introdução

- Projeto PeSO (PUCRS e HP Brazil)
  - ***Pe****squisa em* ***S****istemas* ***O****peracionais escaláveis*
- Escalabilidade de SOs em máquinas NUMA
- Uso de *benchmarks* – gargalos encontrados
- Novas soluções
  - Escalonamento: balanceamento de carga
  - Gerência de memória: migração de páginas

Escalonador do Linux

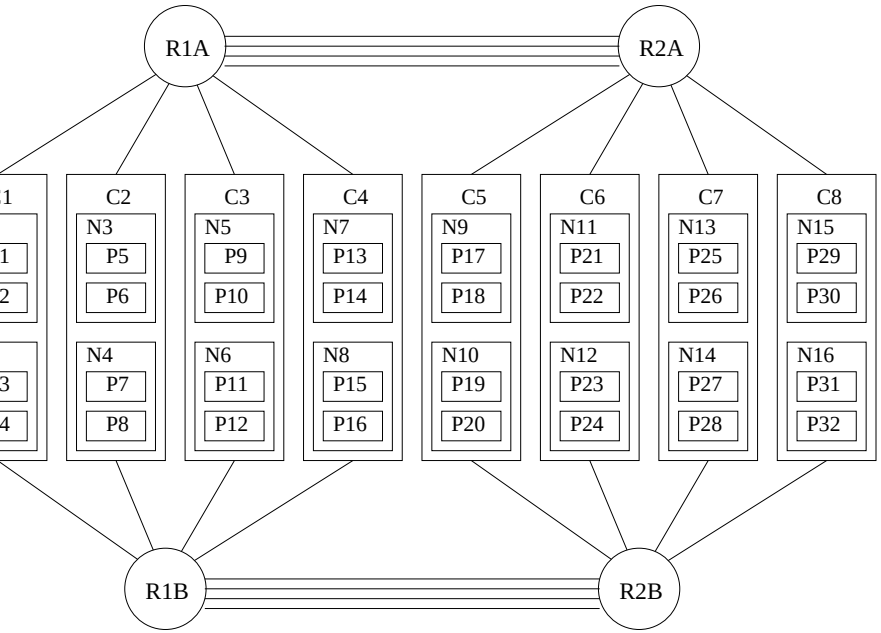
- O(1), desde *kernel* 2.5
- Uma fila de processos por processador (*runqueue*)
  - Afinidade de processador
- Filas de prioridade dinâmicas
- Fatia de tempo dinâmica
- Processos *I/O-bound, CPU-bound*
- *Runqueue* tem dois *arrays*: ativos e expirados

Escalonador do Linux

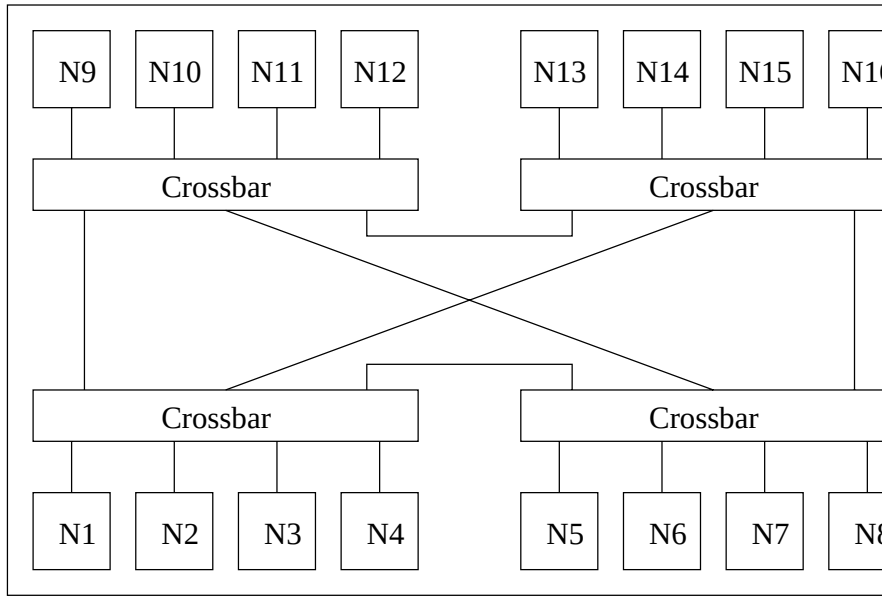


Balancemento de carga: NUMA

Dois exemplos



SGI Altix 3000 Server
(6 níveis de acesso à memória)



HP Integrity Superdome (Orca)
(3 níveis de acesso à memória)

Linux - Balanceamento de carga

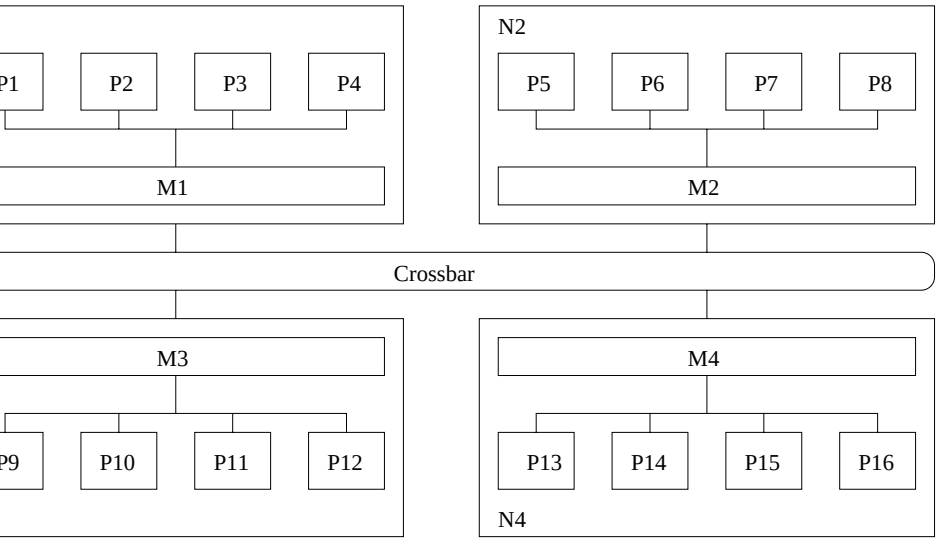
- Sobrecarga do processador
- Para máquinas UMA balanceamento é simples
- Em máquinas NUMA é mais complicado...
- *sched_domain* - representa a topologia da máquina
- SLIT – *System Locality Information Table*
  - ACPI – *Advanced Configuration and Power Interface*

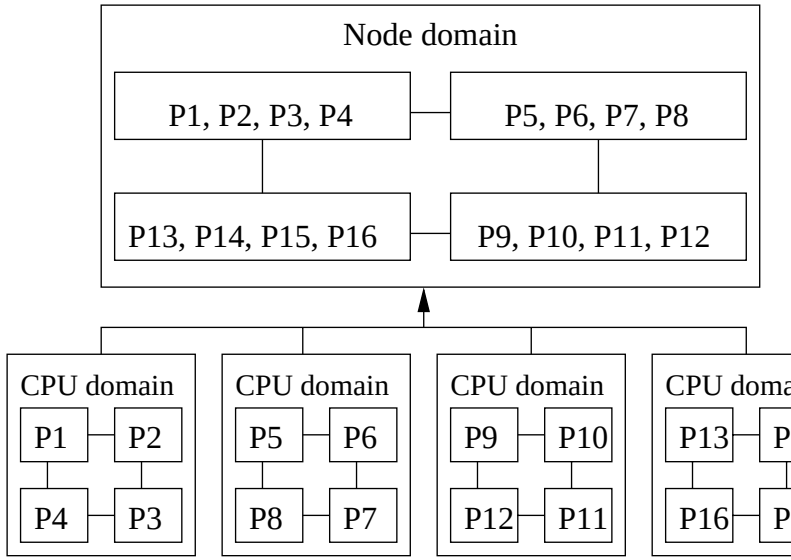
Balanceamento de carga: Linux

- Periodicamente – eventos
- Processos
  - Da fila de processos expirados
  - Processos com prioridades mais alta
- Processos não devem
  - Ter afinidade de processador
  - ser *cache-hot*
- Nova estrutura de domínios de escalonamento

Linux – domínios de escalonamento

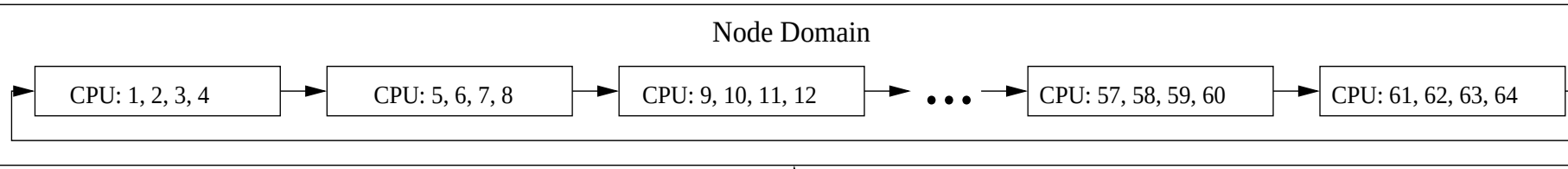
- Exemplo de máquina com dois níveis
- HP Integrity Superdome (Olympia)

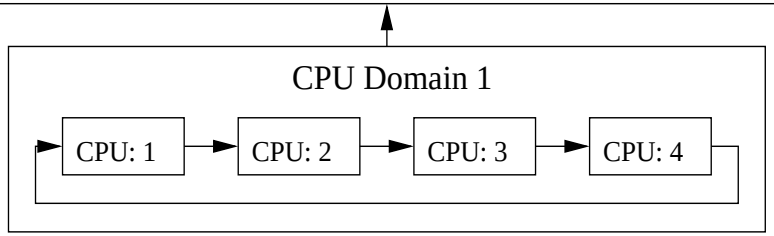


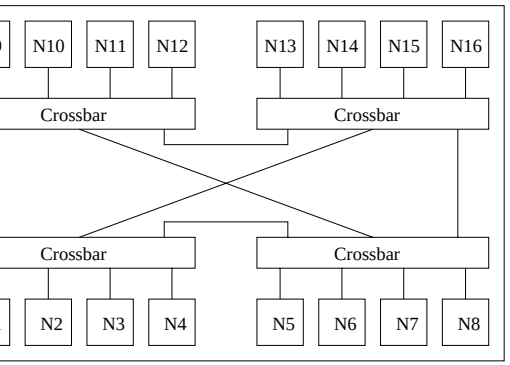


Linux – domínios de escalonamento

- Máquinas com três ou mais níveis - problema
- Linux construía somente estrutura com dois níveis
  - Três se mais de 16 nodos





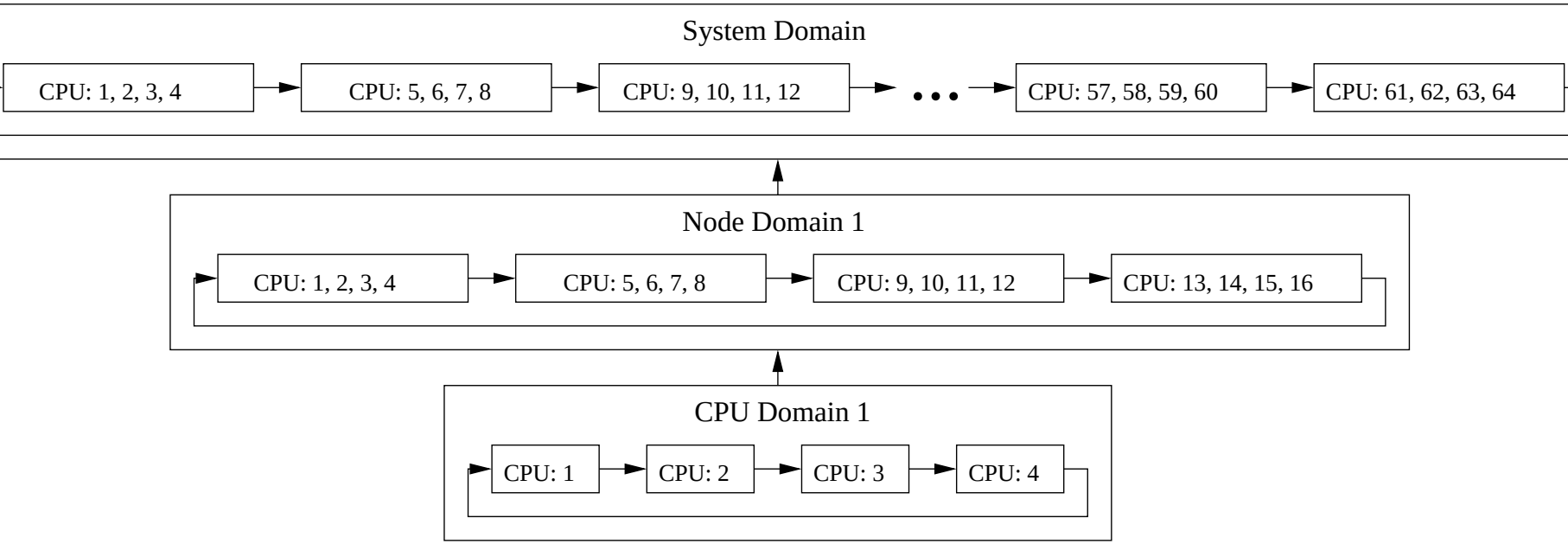


Não era correto !!

Exemplo de tabela SLIT (HP Superdome)

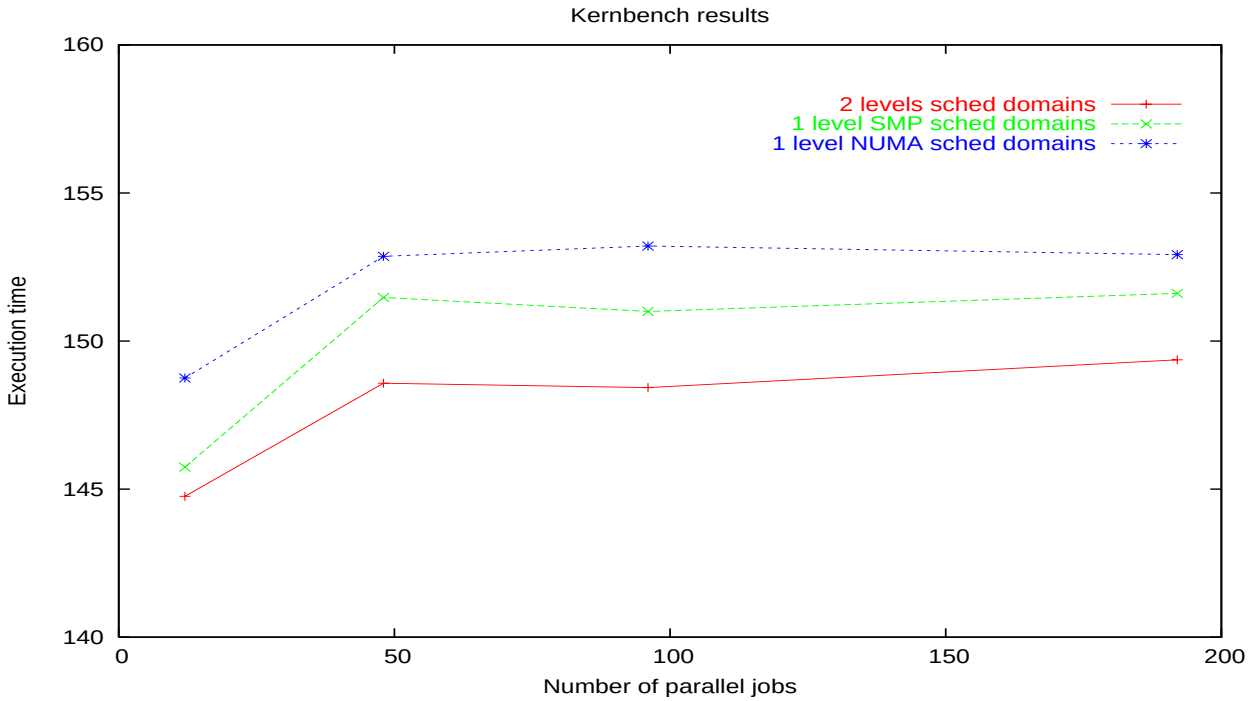
	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	N15	N16
N1	10	17	17	17	29	29	29	29	29	29	29	29	29	29	29	29
N2	17	10	17	17	29	29	29	29	29	29	29	29	29	29	29	29
N3	17	17	10	17	29	29	29	29	29	29	29	29	29	29	29	29
N4	17	17	17	10	29	29	29	29	29	29	29	29	29	29	29	29
N5	29	29	29	29	10	17	17	17	29	29	29	29	29	29	29	29
N6	29	29	29	29	17	10	17	17	29	29	29	29	29	29	29	29
N7	29	29	29	29	17	17	10	17	29	29	29	29	29	29	29	29
N8	29	29	29	29	17	17	17	10	29	29	29	29	29	29	29	29
N9	29	29	29	29	29	29	29	29	10	17	17	17	29	29	29	29
N10	29	29	29	29	29	29	29	29	17	10	17	17	29	29	29	29
N11	29	29	29	29	29	29	29	29	17	17	10	17	29	29	29	29
N12	29	29	29	29	29	29	29	29	17	17	17	10	29	29	29	29
N13	29	29	29	29	29	29	29	29	29	29	29	29	10	17	17	17
N14	29	29	29	29	29	29	29	29	29	29	29	29	17	10	17	17
N15	29	29	29	29	29	29	29	29	29	29	29	29	17	17	10	17
N16	29	29	29	29	29	29	29	29	29	29	29	29	17	17	17	10

Linux – Estrutura correta (Orca)



Resultados - Kernbench

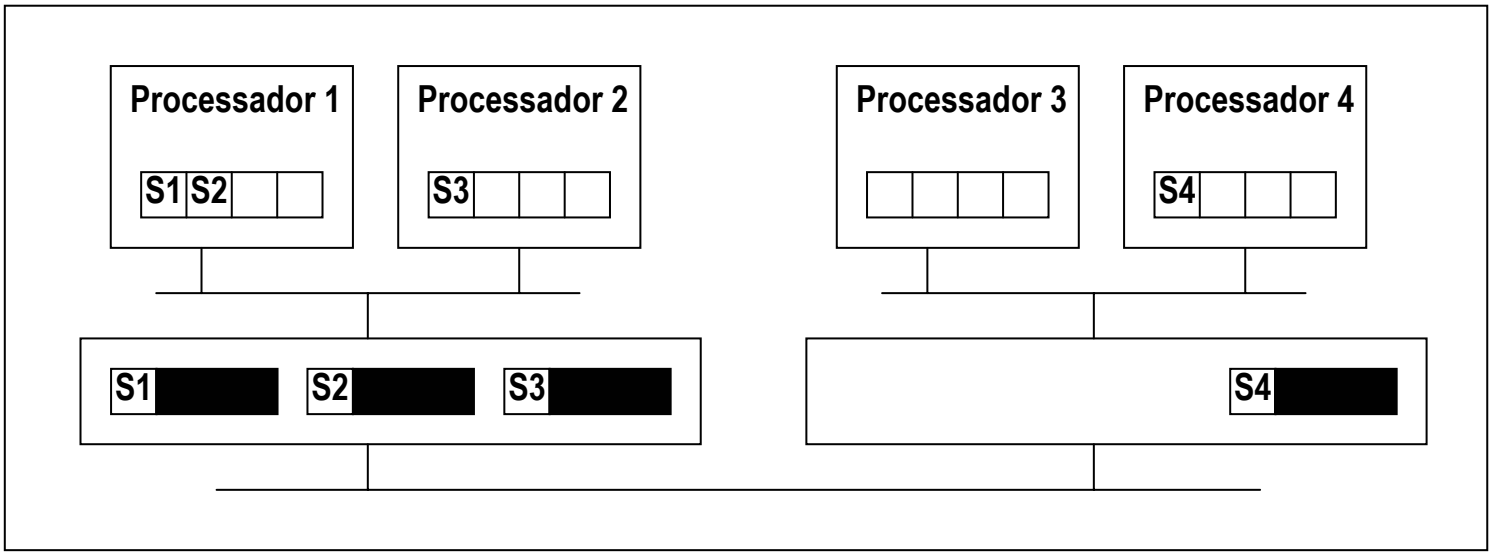
- Três configurações
  - Domínios de escalonamento com dois níveis
  - 1 nível com todos processadores no domínio de CPU (SMP)
  - 1 nível com todos processadores no domínio de Nodo (NUMA)



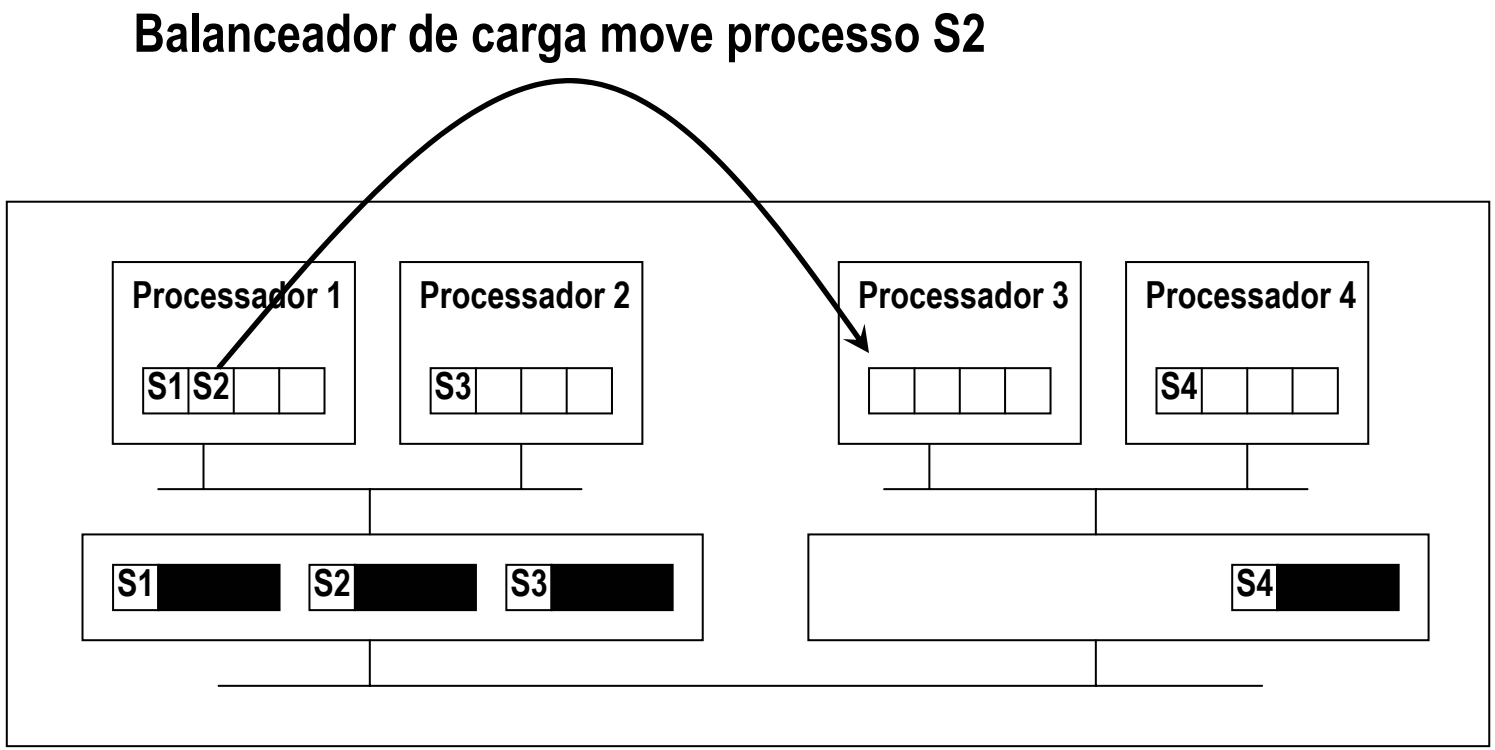
Migração de páginas

- Linux não implementa migração de páginas
- Somente quando ocorre falta de página (*page fault*)
- Acesso à memória fica mais lento
- Estratégias:
  - Migrar todas páginas assim que processo migrar
  - Migrar páginas conforme são acessadas

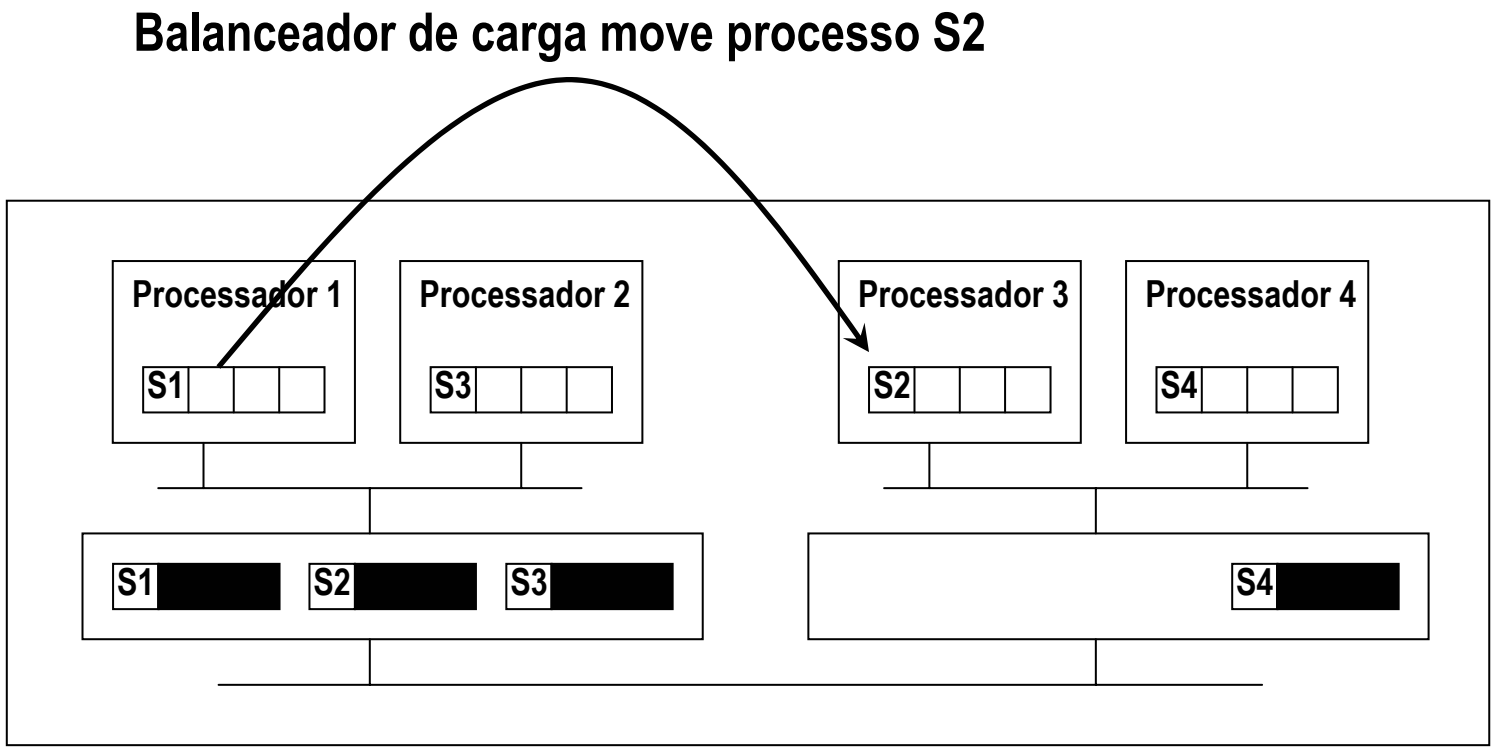
Migrar todas páginas



Migrar todas páginas

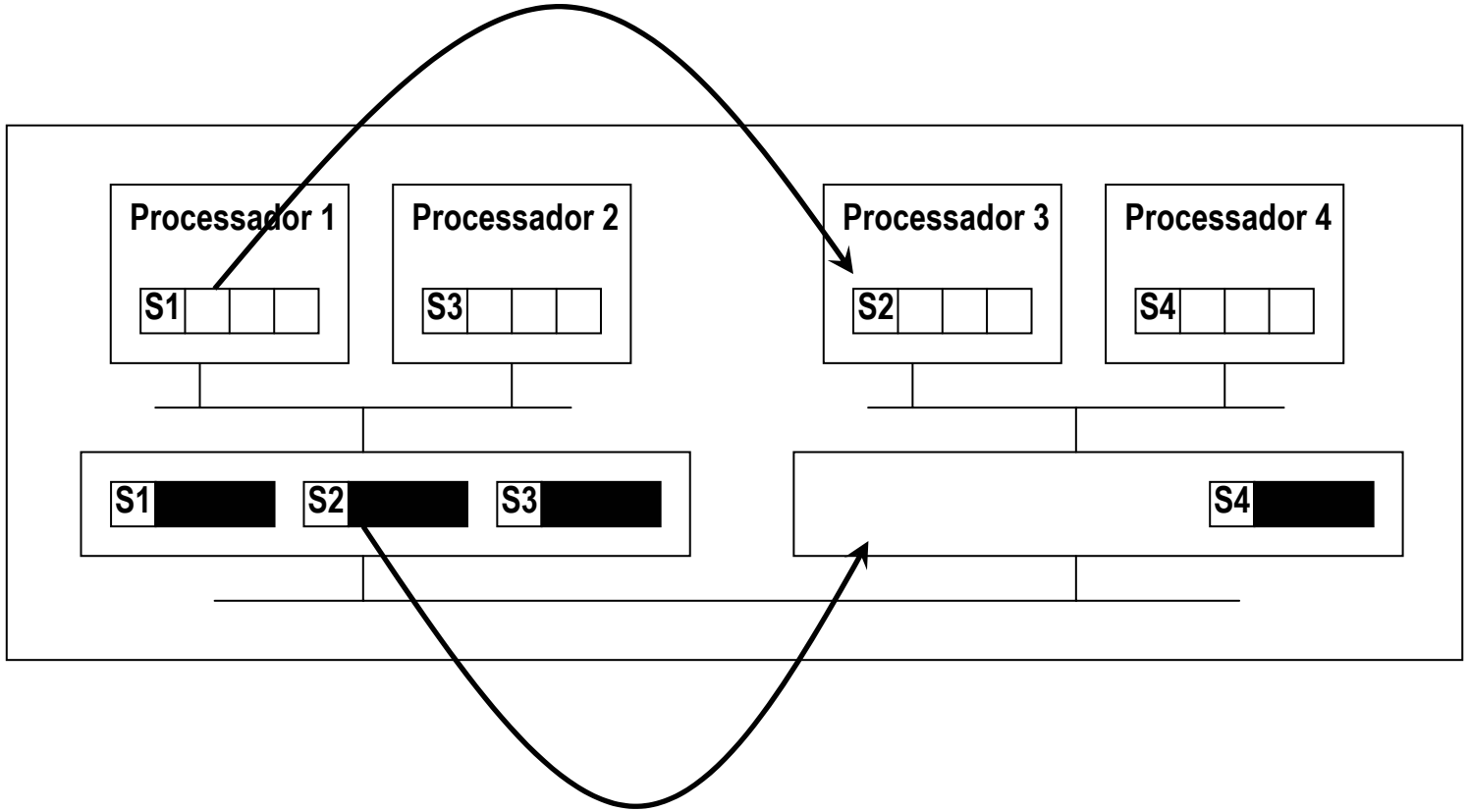


Migrar todas páginas



Migrar todas páginas

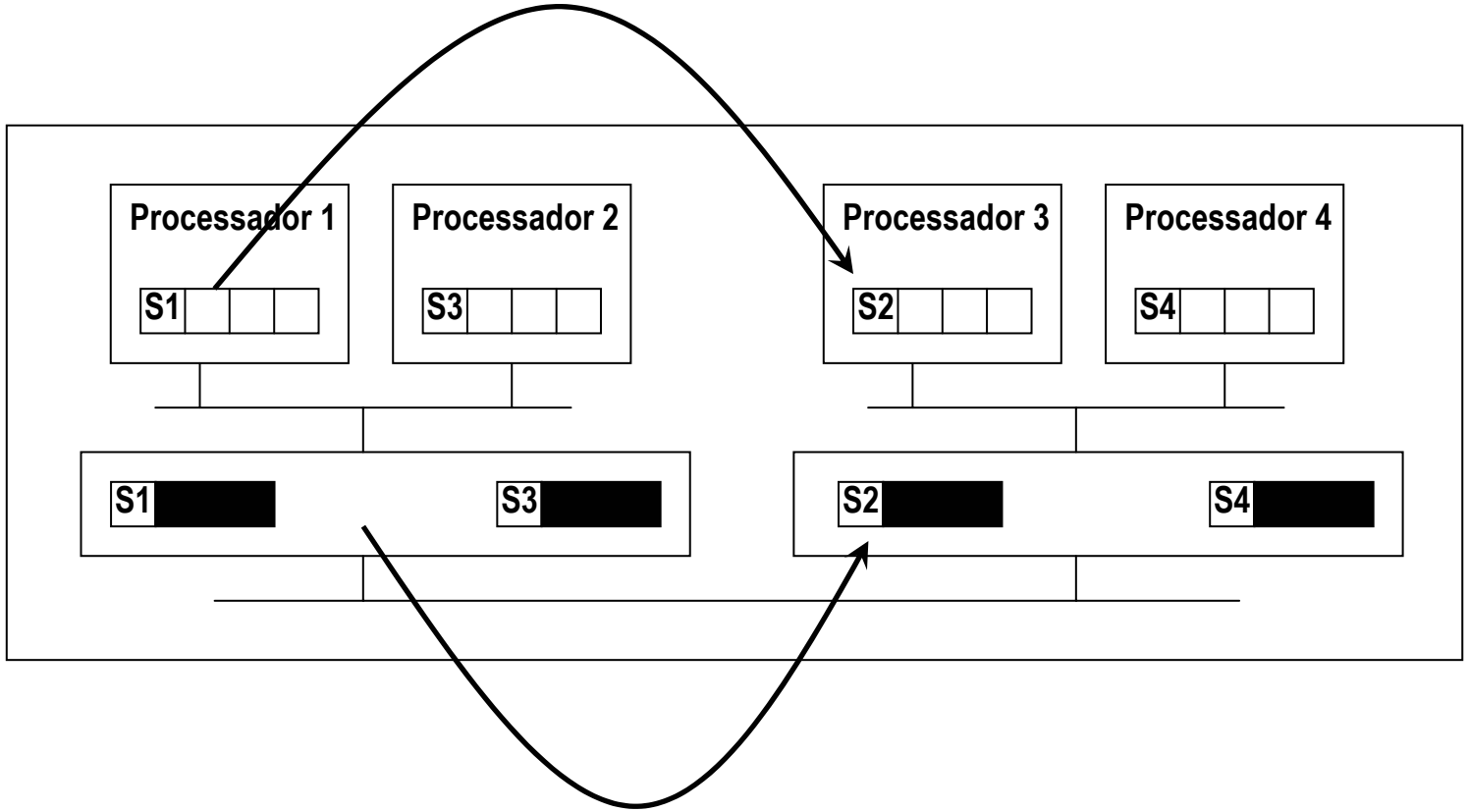
Balanceador de carga move processo S2



Depois move espaço de endereçamento do processo S2

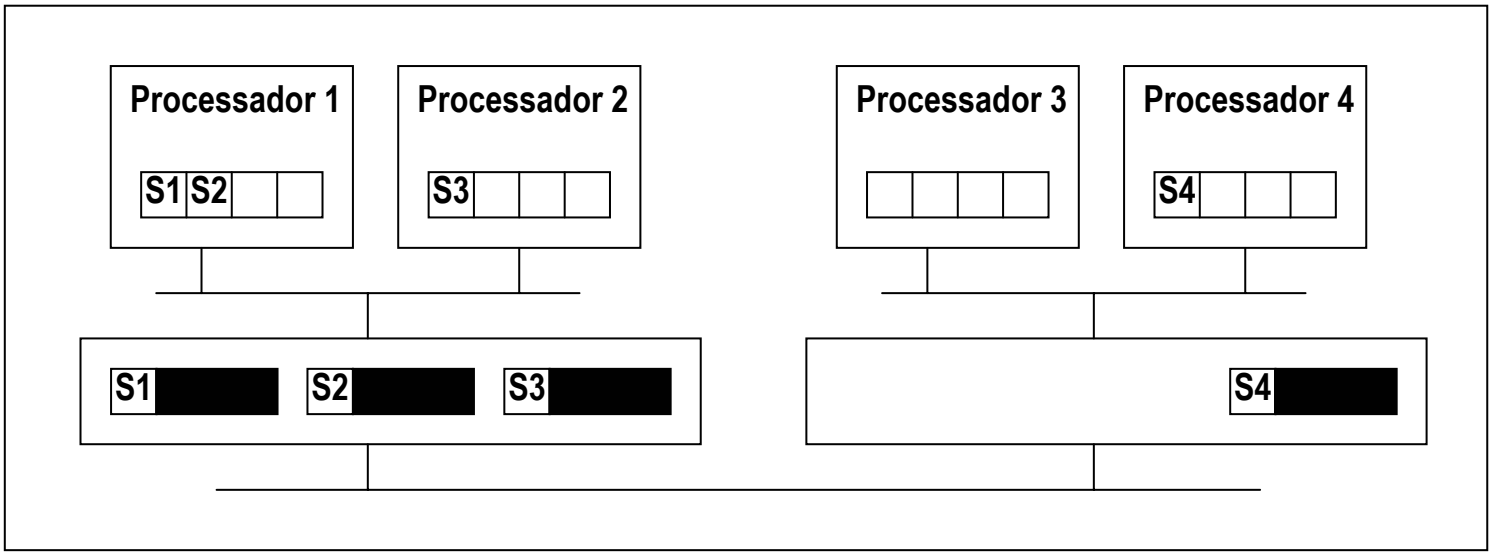
Migrar todas páginas

Balanceador de carga move processo S2

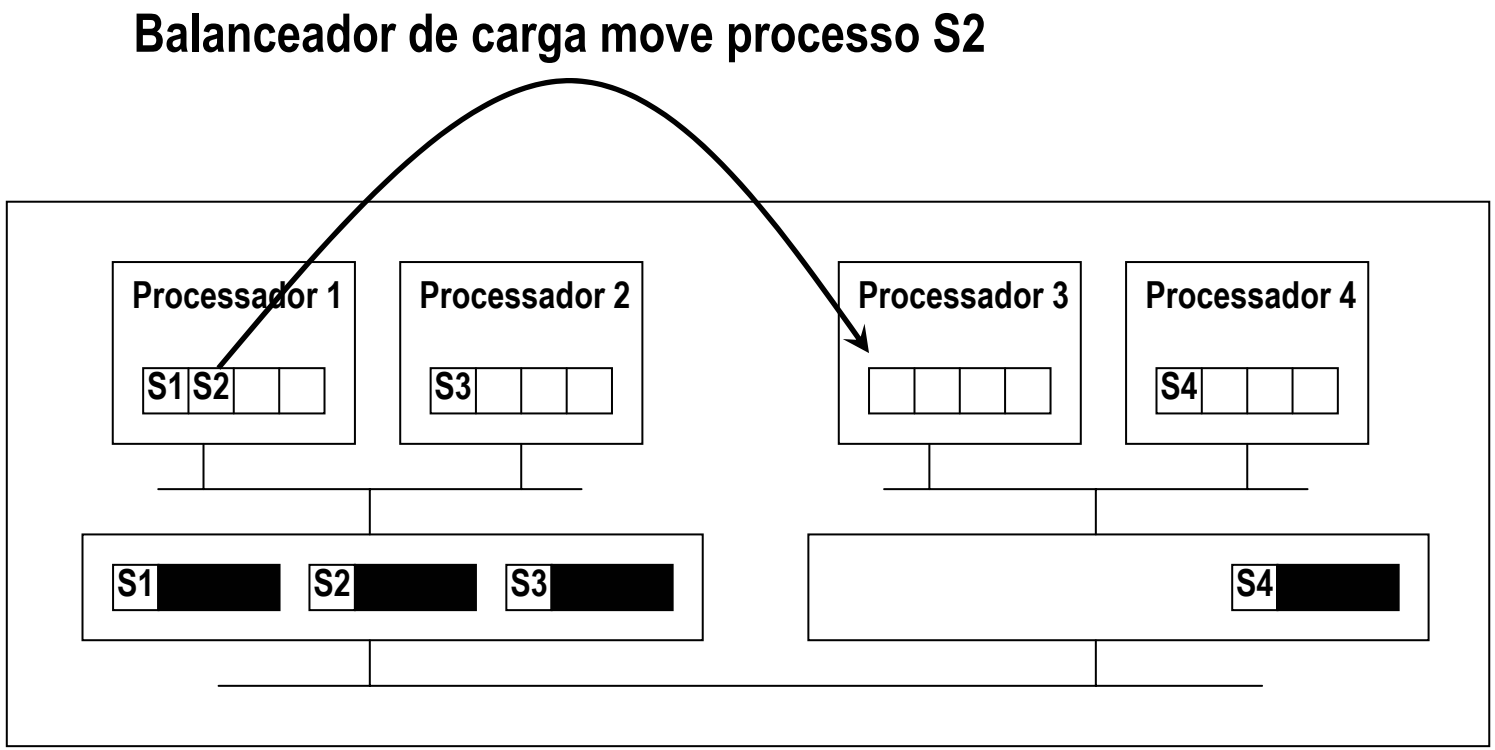


Depois move espaço de endereçamento do processo S2

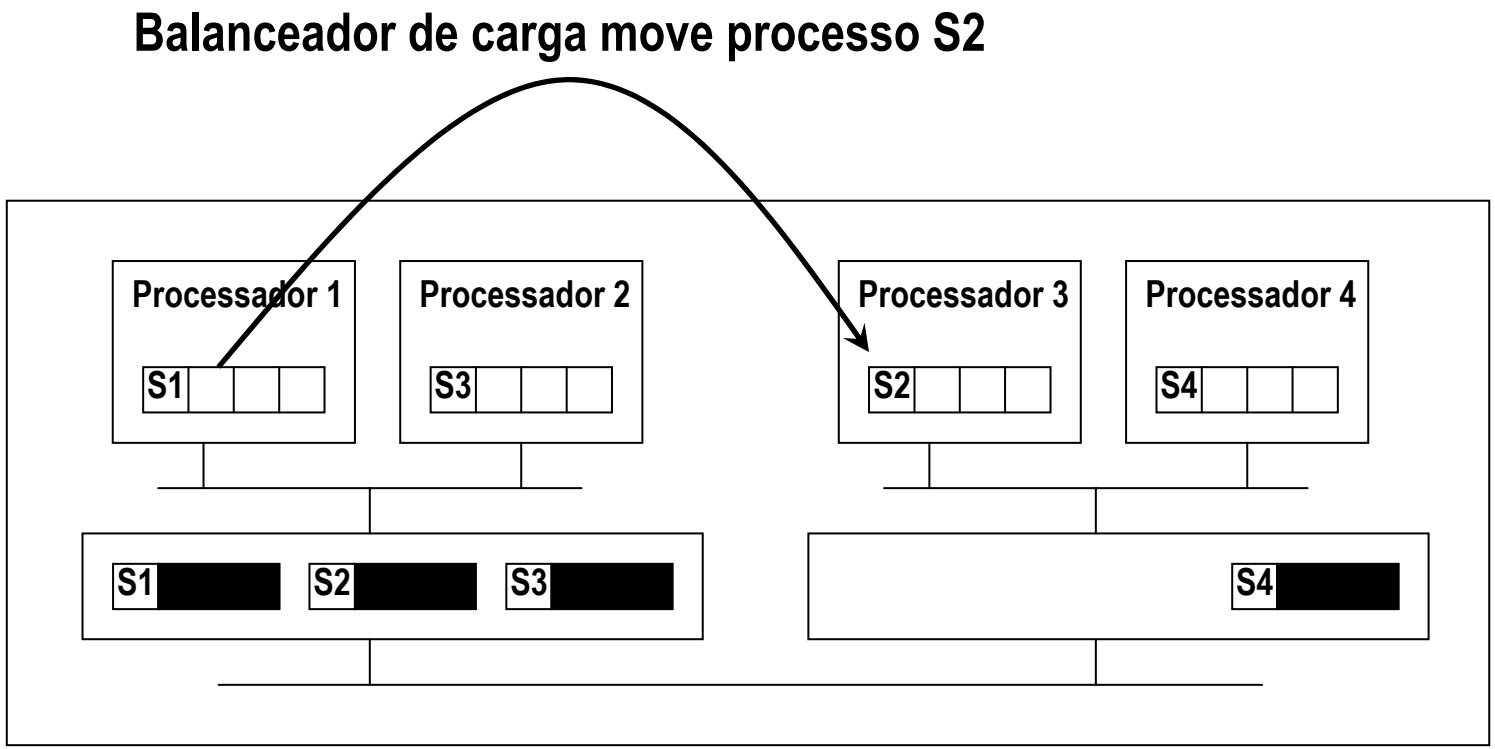
Migrar sob demanda



Migrar sob demanda

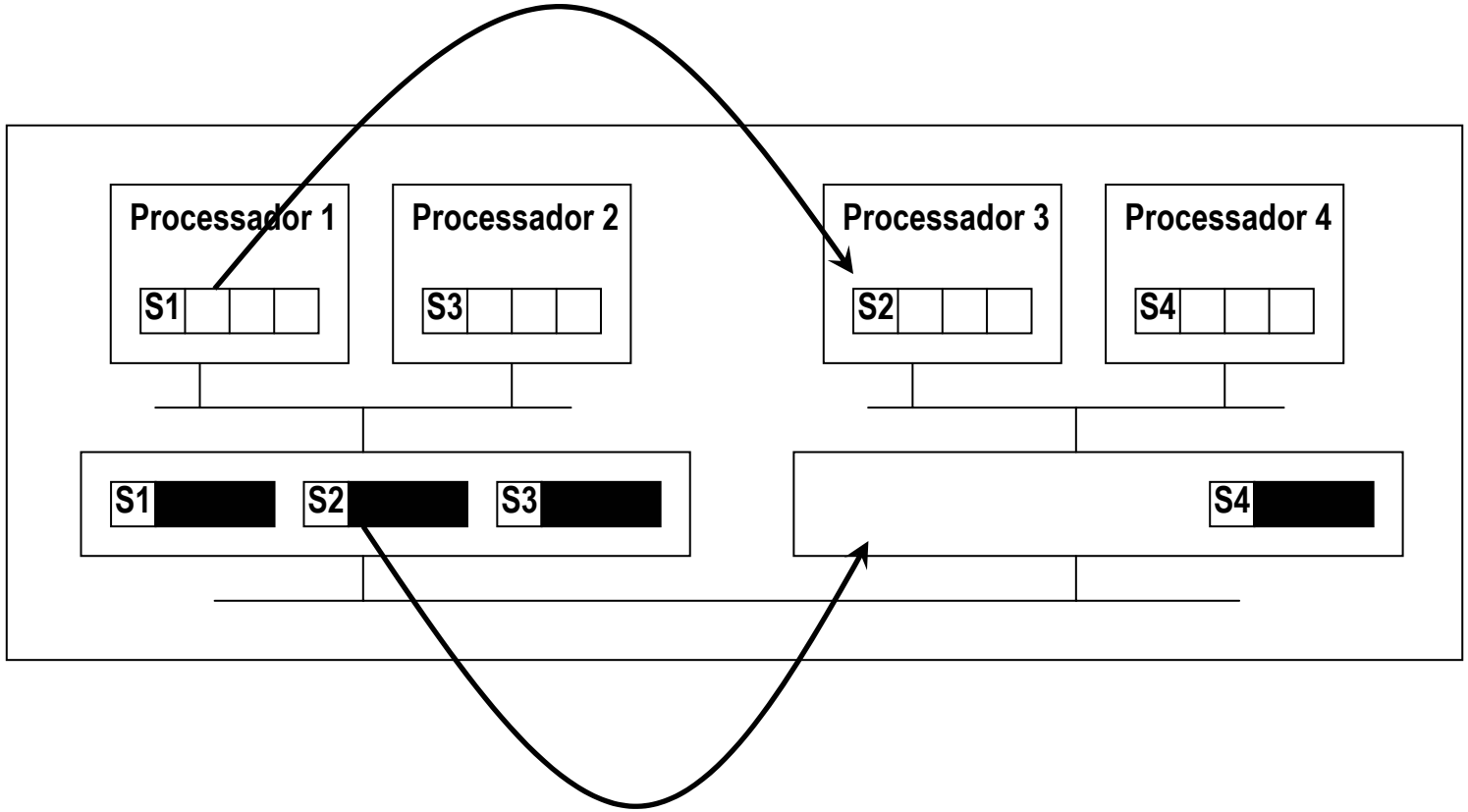


Migrar sob demanda



Migrar sob demanda

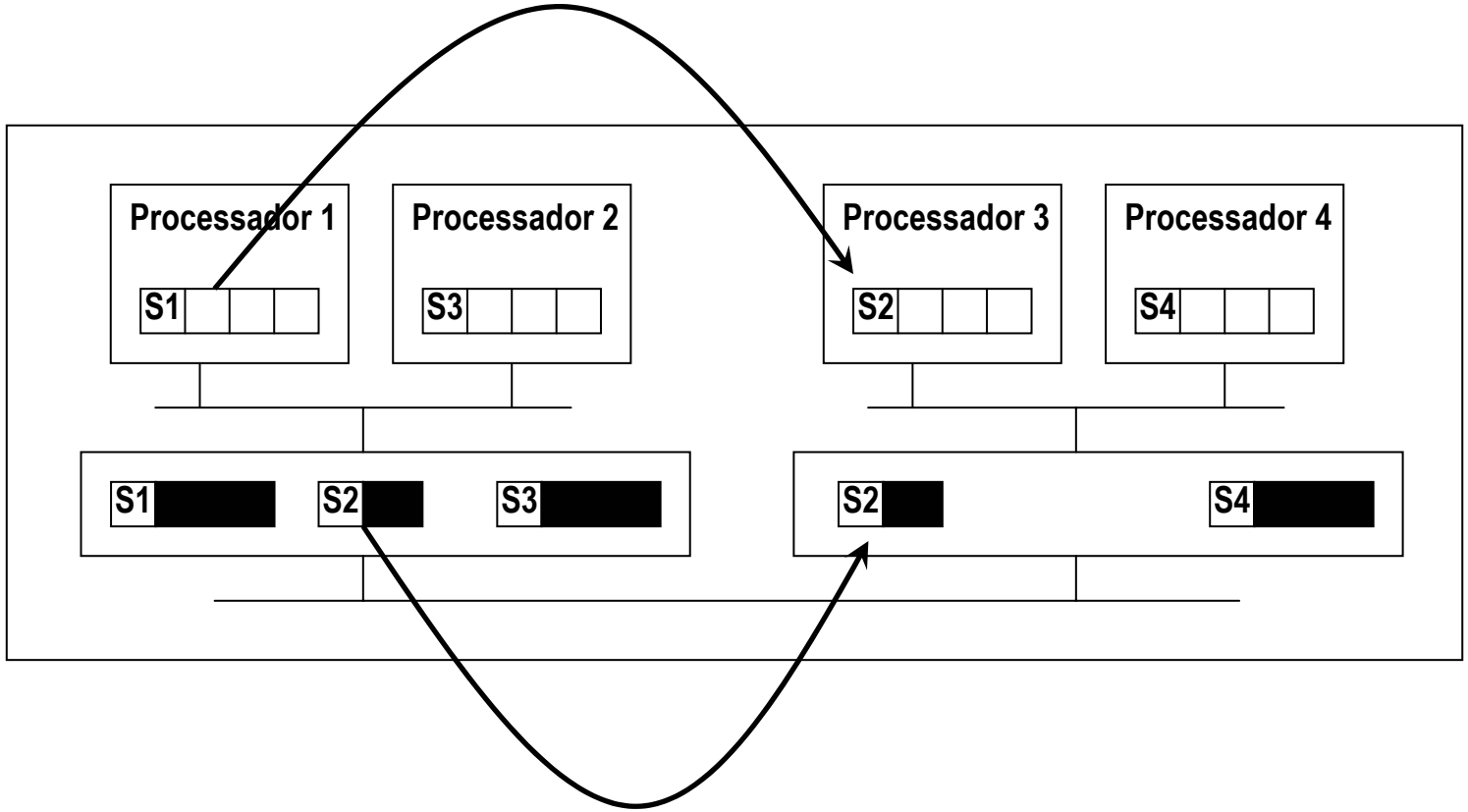
Balanceador de carga move processo S2



Move parte do espaço de endereçamento do processo S2, quando acessado

Migrar todas páginas

Balanceador de carga move processo S2



Move parte do espaço de endereçamento do processo S2, quando acessado

Resultados da simulação

- Comparação com a estratégia atual do Linux
- Percentual de melhora

Tx acesso memória/ Tx acesso rep. memória	20%	40%	60%
20%	1,9%	3,6%	5,1%
40%	3,9%	7,1%	9,9%
60%	5,5%	10,4%	13,7%
80%	7,1%	12,8%	17,5%

Conclusão

- Nova hierarquia para balanceamento de carga
- *Patch* disponível para Linux 2.6.14
  - http://www.inf.pucrs.br/peso
- Novo esquema de migração de páginas
  - Sendo desenvolvido em cooperação com HP Labs
- Melhora do desempenho do Linux para máquinas NUMA