ci1316/ci316 /ci1009 - PROGRAMAÇÃO PARALELA -
1o semestre de 2023
por W.Zola/UFPR

Algoritmos para Broadcast com MPI – motivação e idéias iniciais

---

Segundo um estudo [Laguna et al. 2019] feito sobre o uso do MPI em aplicações de alto desempenho, as operações coletivas representam 99% das funções utilizadas.

O Broadcast é uma operação extremamente importante, sendo a terceira mais utilizada das funções coletivas.

Logo, a sua implementação tem bastante impacto no desempenho de um grande número de aplicações

A operação broadcast do Open MPI é implementada com diferentes algoritmos. De acordo com os parâmetros de cada operação, a biblioteca escolhe automaticamente e executa uma dessas versões.

(continua...)

Em algumas situações, o MPI opta pelo uso da Split Binary Tree.

O algoritmo, apresentado na Figura 1:

Consiste em comunicar as mensagens respeitando a topologia de uma árvore binária.

Divide a mensagem que será comunicada no broadcast, enviando metade dos segmentos para cada sub-árvore.

Faz a "troca de metades" entre nodos simétricos na topologia, em uma etapa final.

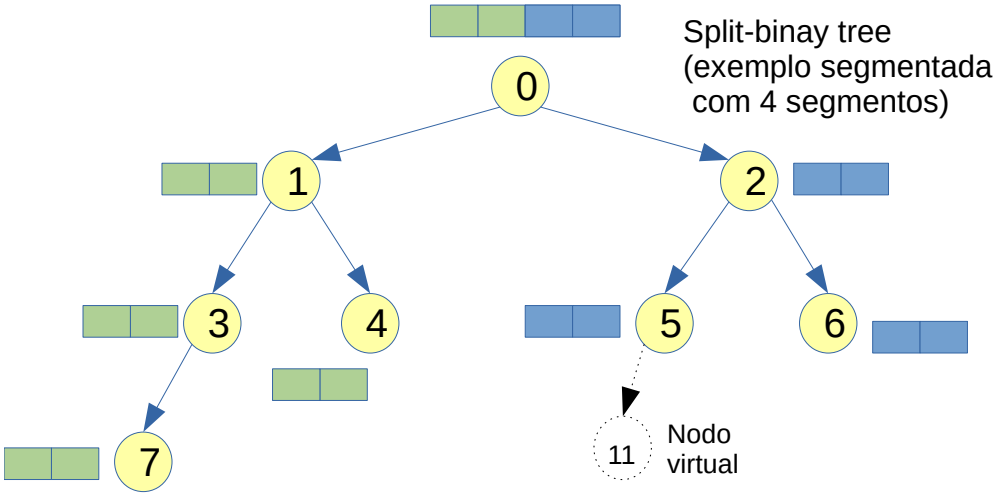


figura da Split Binary Tree

Para o exemplo acima, os nodos "simétricos" que fazem a troca (ou cópia) de metades na etapa final são:

1 com 2

3 com 5

4 com 6

7 com 11
(como 11 não existe, sua "funcao" de troca pode ser assumida por seu pai, o nodo 5, nesse caso 7 troca com 5)

A implementação atual dessa função no Open MPI escolhe seus algoritmos implementados nativamente para executar broadcast seguindo o seguinte critério:
- para 4KB e 16KB, a Split Binary Tree é a opção utilizada.
- para mensagens com os tamanhos acima, o número de processos nao influencia a escolha.

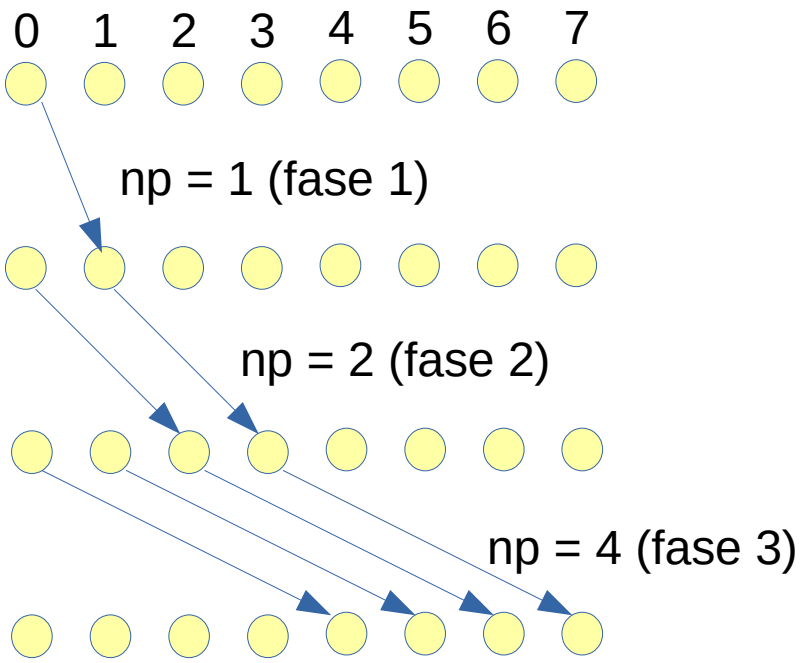
(continua...)

Idéias para o trabalho 2 (my_Bcast_rb)

Neste trabalho vamos implementar um algoritmo de broadcast em MPI e comparar com a versao usada no MPI nativo.

Testaremos com os tamanhos de mensagens 4KB e 16KB, pois nesse caso a Split Binary Tree é a opção utilizada internamente pelo MPI.

No entanto, em vez de reimplementar a Split Binary Tree vamos usar um outro tipo de arvore mostrado abaixo na fig 2, para broadcast com 8 nodos.



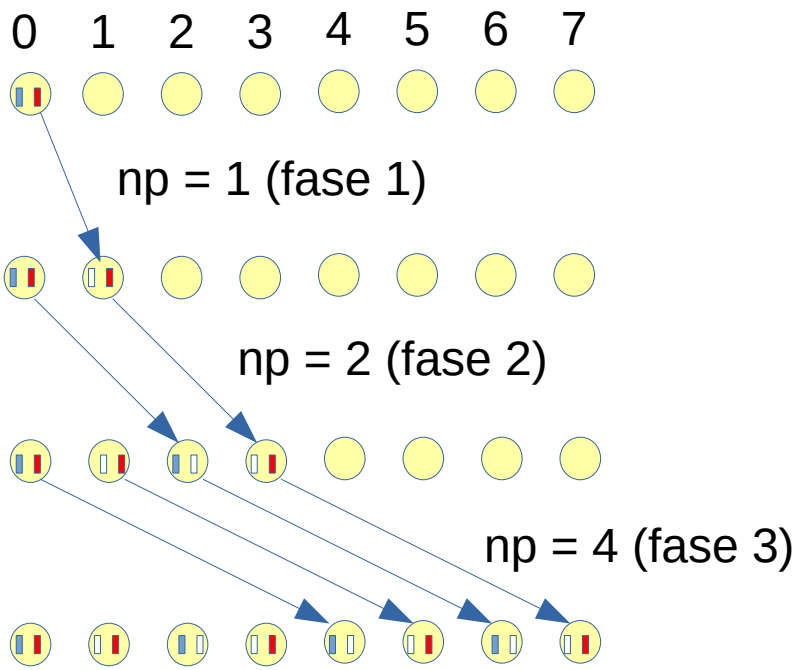
Idéias para o trabalho 2 (my_Bcast_rb) CONTINUACAO

Além de distribuir as mensagens conforme a nossa nova topologia de arvore:

- também faremos a "quebra de mensagens" em metades como feito pela Split Binary Tree.
- também faremos a "troca de metades" em etapa final.

A figura a 3 mostra como de maneira mais complleta, com faremos nosssa implementaçao.

A descricao completa de como poderemos implementar será apresentada em outro texto.



Note que nesse ponto a maioria dos nodos tem apenas uma metade da mensagem.

Falta a etapa final:
- Basta cada nodo par trocar sua mensagem com o nodo impar vizinho.