

## 20 - MPI em cluster

SuperComputação 2019/2

Igor Montagner, Luciano Soares

Objetivos de aprendizagem:

1. Configurar um cluster MPI e compartilhar dados entre as máquinas
2. Compilar e executar os primeiros programas com MPI,
3. Realizar testes de latência de comunicação local e via rede.

Neste roteiro iremos configurar um conjunto de PCs linux para funcionar como um cluster MPI e rodar nossos primeiros programas. Junto com este roteiro você está recebendo um cartão com uma lista de IPs e chaves que funcionam em todos eles.

### Configuração inicial

O primeiro passo para configurar um cluster MPI é permitir a autenticação sem senha entre o nó mestre e os restantes. A lista de máquinas contém uma máquina principal para cada aluno. Este guia foi parcialmente baseado no tutorial [Running an MPI Cluster within a LAN](#) escrito por Dwaraka Nath.

### Autenticação

A primeira etapa de configuração do cluster é criar usuários em cada máquina e permitir `ssh` sem senha entre elas.

O processo abaixo deverá ser feito em 3 máquinas do cluster: sua principal mais duas a sua escolha.

1. Faça o login usando a chave disponibilizada no blackboard.
2. Crie um usuário com seu nome;
3. Crie uma chave criptográfica usando `ssh-keygen`;

Com todos os usuários criados, logue em cada máquina e copie sua chave pública para o `~/.ssh/authorized_keys` das outras duas máquinas e faça um login de teste.

**Importante:** estas instruções permitem o ssh sem senha de uma máquinas para todas as outras. É importante testar a conexão antes de executar os processos usando MPI.

Estes passos são necessários para criar um cluster seguro em clouds públicas.

### Instalação de software

Instale os pacotes necessários, compile e rode o exemplo `hello.cpp` usado na aula passada.

### Compartilhamento de dados e instalação de software

Para compartilhar dados entre nossas máquinas iremos configurar uma partição *NFS* (*Network File System*). Todo arquivo colocado na pasta *NFS* estará disponível em todas as máquinas do nosso cluster e é nele que estará nosso executável. No caso do projeto 3, podemos usar este espaço também para salvar dados de progresso da análise de cada site.

Siga [este guia](#) para configurar o compartilhamento de arquivos. A máquina **master** dele é sua máquina principal e a parte **client** deve ser replicada para as duas máquinas que você escolheu.

Com tudo configurado, copie os arquivos `hello_cluster.cpp`, `mpi_bandwidth.c` e `mpi_latency.c` para a área compartilhada e passe para a próxima seção.

## Primeiros usos do cluster

Vamos então rodar o programa *hello\_cluster.cpp* em várias máquinas. Para isto é necessário criar um arquivo *hostfiles* em que cada linha descreve o ip de uma máquina remota e quantos processos ela pode rodar. Neste roteiro iremos rodar todos os exemplos a partir da nossa máquina principal na AWS. Criem um arquivo com o seguinte formato e os IPs de todas as máquinas do nosso cluster. Note que estamos enviando dois processos para cada máquina do cluster

```
ip_maquina1
ip_maquina2
localhost
```

E executem o hello com a seguinte linha de comando (supondo que o arquivo acima seja chamado *hosts*):

```
> $ mpiexec -n 3 -hostfile hosts ./hello_cluster
```

## Custos de comunicação

Um dos grandes diferenciais de utilizar um cluster ao invés de uma máquina é o custo de comunicação. Ao executar MPI usando processos locais o custo de comunicação é muito menor que quando processos são executados em máquinas remotas. Iremos quantificar este custo usando duas medidas.

1. latência: tempo que uma mensagem demora para ser entregue (não importa o tamanho da mensagem)
2. largura de banda: quantidade de informação que pode ser enviada por segundo (levando em conta já o custo de comunicação).

Os programas *mpi\_latency.c* e *mpi\_bandwidth.c* calculam estas medidas no cluster. Compile e execute ambos os programas em duas situações: rodando com três processos localmente e rodando com processos em suas três máquinas. Compare os resultados das duas execuções abaixo.

### Resultados locais

### Resultados cluster (3 máquinas)

### Comentários