

Uma introdução ao Apache Hama

O modelo BSP nas nuvens

Thiago Kenji Okada¹

¹Departamento de Ciências da Computação (DCC)
Instituto de Matemática e Estatística (IME)
Universidade de São Paulo (USP)
MAC5742 – Computação Paralela e Distribuída

26 de Junho de 2015



- 1 Introdução
 - O que é o Apache Hama?
 - O que é o modelo BSP?
- 2 Apache Hama e o modelo BSP
 - Mais detalhes sobre o Apache Hama
 - Escrevendo aplicações BSP no Apache Hama
- 3 Possíveis aplicações para o Apache Hama
- 4 Conclusões



Roteiro

- 1 Introdução
 - O que é o Apache Hama?
 - O que é o modelo BSP?
- 2 Apache Hama e o modelo BSP
 - Mais detalhes sobre o Apache Hama
 - Escrevendo aplicações BSP no Apache Hama
- 3 Possíveis aplicações para o Apache Hama
- 4 Conclusões



O que é o Apache Hama?

- O Apache Hama é um *framework* de propósito geral, escrito em Java, baseado no modelo *Bulk Synchronous Parallel* (BSP);
- Usa a plataforma do Apache Hadoop;
- Desde 2012, é um projetos de alta prioridade da Fundação Apache.

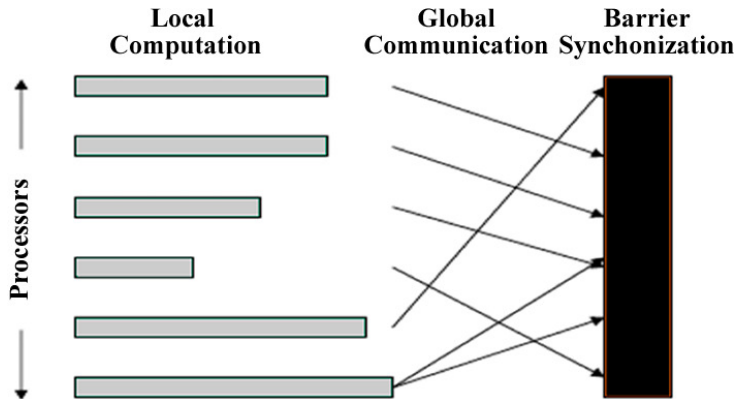


O que é o modelo BSP?

- Modelo de programação paralela *Bulk Synchronous Parallel* proposto em 1990 por Valiant;
- Composto de um modelo computacional e um modelo de algoritmo:
 - Um **computador BSP** é composto de:
 - Um conjunto de processadores, cada um com sua memória;
 - Um rede, que permite troca de mensagens;
 - Um mecanismo de sincronização entre processadores.
 - Um **algoritmo BSP** é composto de uma sequência de *super-passos*:
 - Processamento local;
 - Comunicação entre processos;
 - Barreira de sincronismo.



Super-passo no modelo BSP

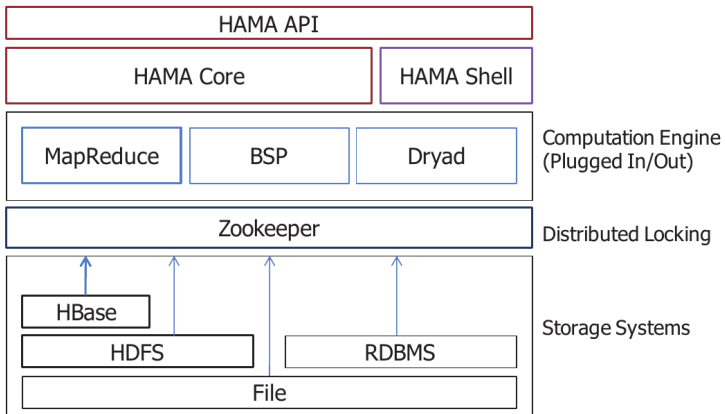


Roteiro

- 1 Introdução
 - O que é o Apache Hama?
 - O que é o modelo BSP?
- 2 Apache Hama e o modelo BSP
 - Mais detalhes sobre o Apache Hama
 - Escrevendo aplicações BSP no Apache Hama
- 3 Possíveis aplicações para o Apache Hama
- 4 Conclusões



Arquitetura geral do Apache Hama



Seo et al. (2010) [1]



Objetivos do Apache Hama

De acordo com Seo et al. (2010) [1], as principais contribuições do Apache Hama foram:

- **Compatibilidade:** usa funcionalidades já presentes no Hadoop;
- **Escalabilidade:** usa as mesmas interfaces do Hadoop;
- **Flexibilidade:** diversos motores de processamento diferentes, e é possível implementar novos;
- **Aplicabilidade:** útil para problemas envolvendo grafos e matrizes.



Estrutura básica de uma aplicação do Apache Hama

- Começamos extendendo o método `org.apache.hama.bsp.BSP`;
- Sobrescrevemos o método `bsp()` dessa classe;
- Usamos uma das diversas primitivas de comunicação disponíveis no Hama para troca de mensagens;
- Finalizamos com uma chamada ao método `sync()` da classe `BSPPeer`.



Exemplo de aplicação no Apache Hama

@Override

```
public void bsp(BSPPeer<NullWritable, NullWritable, Text, Text,  
    Text> peer) throws IOException, SyncException,  
    InterruptedException {  
    int i;  
    for (i = 0; i < 100; ++i);  
  
    for (String peerName : peer.getAllPeerNames()) {  
        peer.send(peerName, new Text("Hello from " +  
            peer.getPeerName() + ", i=" + i));  
    }  
  
    peer.sync();  
  
    System.out.println(peer.getCurrentMessage());  
}
```



Primitivas de comunicação

Função	Descrição
<code>send(String peerName, BSPMessage msg)</code>	Envia uma mensagem para outro nó.
<code>getCurrentMessage()</code>	Recebe uma mensagem da fila.
<code>getNumCurrentMessages()</code>	Obtém o número de mensagens atuais da fila.
<code>sync()</code>	Inicia a barreira de sincronismo.
<code>getPeerName()</code>	Obtém o nome do nó da tarefa atual.
<code>getPeerName(int index)</code>	Obtém o nome do nó de índice n.
<code>getNumPeers()</code>	Obtém o número de nós.
<code>getAllPeerNames()</code>	Obtém o nome de todos os nós, incluindo o atual, em ordem crescente.



Roteiro

- 1 Introdução
 - O que é o Apache Hama?
 - O que é o modelo BSP?
- 2 Apache Hama e o modelo BSP
 - Mais detalhes sobre o Apache Hama
 - Escrevendo aplicações BSP no Apache Hama
- 3 Possíveis aplicações para o Apache Hama
- 4 Conclusões



Possíveis aplicações para o Apache Hama

- Multiplicação de matrizes;
- Solução de sistemas lineares;
- Aprendizado de máquina;
- Processamento de grandes volumes de dados em geral;
- **Redes sociais.**



Roteiro

- 1 Introdução
 - O que é o Apache Hama?
 - O que é o modelo BSP?
- 2 Apache Hama e o modelo BSP
 - Mais detalhes sobre o Apache Hama
 - Escrevendo aplicações BSP no Apache Hama
- 3 Possíveis aplicações para o Apache Hama
- 4 Conclusões



Conclusões

- Ele soluciona uma classe de problemas que não funcionam bem usando o modelo mais tradicional do Apache Hadoop (o MapReduce).
- O potencial existe, o problema é que não existe muita gente trabalhando nele;
- Concorrência forte com o Apache Giraph e o **Apache Spark**.



Fim da apresentação



Referência



Seo, Sangwon, et al.

Hama: An efficient matrix computation with the mapreduce framework

2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom) (2010): 721–726.

