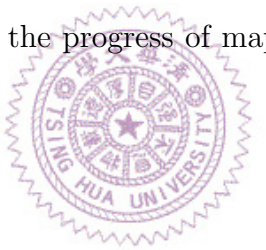
Abstract

As more and more data generated every moment over the internet, the requirement of method to solve new scale problems is getting important. Hadoop mapreduce is one of the most important tools which widely used on solving large scale and rapidly growing problems in today's big data era. Based on traditional mapreduce framework, many researches proposed their strategy to adapt different situations. But to actually full use the resource of hadoop cluster, we still require a new framework to allocate tasks. In this thesis, we develop a resource-aware scheduling strategy to overcome the drawbacks of traditional framework, and propose a mismatch controlling algorithm that coordinates the progress of mapper and reducer to achieve the full usage of resource.



Index Terms: mapreduce,resource,mismatch

中文摘要

隨著網路上日漸增多的資料量，能夠處理新數量級問題的方法也更加重要。在今日的巨量資料時代，Hadoop mapreduce是其中一種被廣泛運用來處理大量、成長快速資料的重要工具。許多的研究在傳統的mapreduce架構上提出他們的策略來適應不同的情況，然而，為了達到真正使用到叢集裡所有資源，我們仍然需要一個新的架構來分配工作。在這篇論文中，我們研發了一個資源控管策略來克服傳統架構上的缺點，並提出進度平衡演算法來調控Mapper與Reducer間的進度不平衡，藉此達到叢集內資源的高利用率。

關鍵詞：mapreduce，資源，不平衡