doi:10.1631/FITEE.1601056

题目：满足 MapReduce 环境下近似处理的时限要求

概要：为了向大数据分析提供实时结果，在现今的生产环境中满足 MapReduce 作业的时限要求是非常关键的。许多研究致力于解决时限要求的问题，目前存在两种代表性的方法。第一种是分配适量资源以在时限前完成整个作业，在时限紧迫或资源受限时，该方法会错过时限；另一种是在时限约束下运行预数据量的样本，该方法能满足时限但无法使数据量最大化。在本文中，我们提出一个时限－导向的任务调度方法来解决上述问题，称为“Dart”。给定具体的时限和可用资源量时，Dart 使用基于历史数据和作业运行状态的迭代估计法准确预测作业完成时间。基于时间预测，Dart 法采用接近－修改算法做出动态调度决策，在满足时限的情况下将可处理数据量最大化并消除掉队任务。同时 Dart 法可有效地避免任务失败和数据倾斜，防止其性能受影响。在包含 64 个虚拟机的集群上使用 OpenCloud 和 Facebook 的工作负载对 Dart 法进行评估。结果表明 Dart 法在时限紧迫和资源受限情况下能有效满足时限并将处理数据量最大化。

关键词：MapReduce；近似作业；时限；任务调度；掉队任务消除