超大文本文件单词频数统计

指导教师：杨刚

2023 年 10 月 11 日

线程是现代主流操作系统用来提高系统并发度、资源利用率以及应用效率的关键概念和重要机制。采用 “分而治之” 的策略，将一个规模较大、处理复杂的问题划分较小的问题，每个小问题利用一个工作线程 (Worker) 单独处理，已经成为现代复杂系统优化设计的有效手段。

现在我们通过一个实例来深化对线程概念的理解。假设你需要统计某个关键字（Key Word）在某超大文本文件（$\geq 1G(2^{20})$Bytes）出现的频数，要求：

问题 1.

分别采用单进程/单线程、多进程和线程池[1]的方式解决此问题。

问题 2.

统计分析问题1在上述若干种方式下的性能（平均完成时间）差异，并给出解释。

问题 3.

请调整多线程实现方案中的线程数，统计在不同线程数条件下平均完成时间，挖掘线程数与完成时间之间的关联并给出解释。

Be active!

请以 PPT 的形式完成报告。我将邀请 $2 \sim 2$ 组同学在课堂上进行交流。

[1]可选。