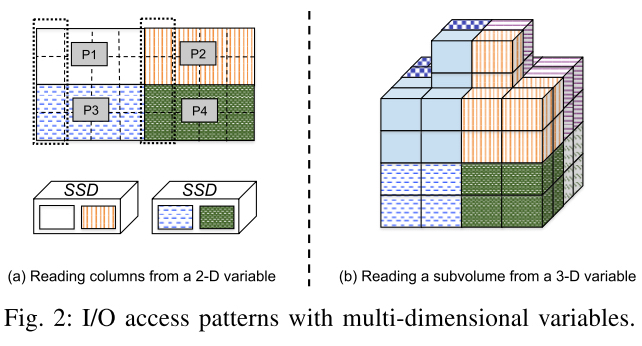
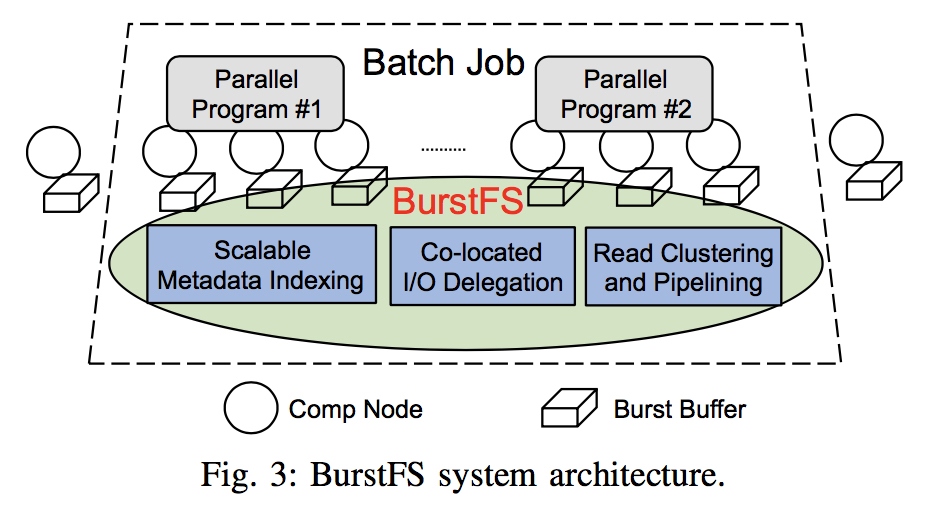
An Ephemeral Burst-Buffer File System for Scientific Applications

摘要：本文设计了一个node-local的BBFS—BurstFS，从而更好得支持科学应用中的N-1写检查点、访问多维数据造成小的不连续读的IO访问特征。

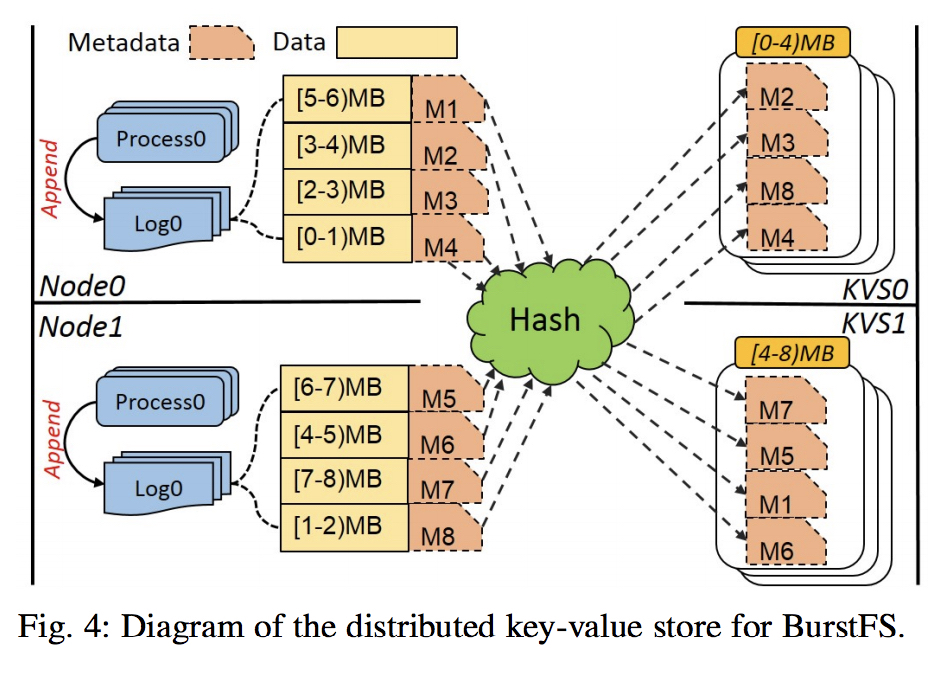
1. 动机
2. 对于node-local的burst buffer系统，N-1的写是一个问题。
3. 在科学应用中由于数据经常是多维的，所以造成访问数据时可能会出现不连续的IO。



1. 设计
2. 整体架构：BurstFS是node-loacl的BBFS。



1. 元数据管理：采用分布式的KV存储（MDHIM），一个片段的数据对应一个元数据。具体分配到哪个机子上通过hash计算后得到。
2. 数据管理：写数据的时候只能写在本地节点，从而让N-1的写扩展性很好。



1. 通信时派代表：1个node上所有进程公用1个delegator进程进行数据请求与应答；其中，client在发送请求的时候会对连续的请求进行合并；server接受请求的时候也会对收到的连续数据的请求进行合并成大请求，并且优先处理大的请求，从而缓解了科学计算中小的不连续读的问题。此外，发送应答数据的时候使用流水线，使得从 SSD中读取数据和发送数据的操作可以并行。

