Dask: Parallel Computation with Blocked algorithms and Task Scheduling

wbg231

January 2023

1 abstract

- dask enables parallel and out of core computation.
- we couple vlocked algorirhsm wirh memory awate task scheudling to achive a parallel out of core Numpy clone
- this scales to modern hardware and large datasets

Introduction

- sci py stuff does not use parallel implementations
- we want to parallelize the scipy code with out needing a full re-write
- dask encodes parallel algorithms using python primitives, and has hte dask.array type a parallel n dimensional array that copies numpy's interface

modern hardware

- hardware has changed a lot in recent years
- most modern cpu's have multiple threads, most modern storage is on an ssd which makes reading information from disk much faster and thus more practice
- these advancements make single machine implementations rival small cluster computation while keeping the ease of working with a single machine

dask graphs

- dask encodes parallel computation in a way that requires low amount of instruction by the developer
- a dask graph is a python dictionary mapping keys or tasks to values
- storing programs in graphs allows for easy task scheduling

specification

- represent computation as a DAG of tasks with data dependincies
- a task is a tuple with a callable first ellement
- tasks are automic unfits of work that can be run by a single worker
- an argument may be either a key present in the dask, a literal, another task,, or a list of arguments

dask arrays

- the dask array submodule uses dask graphs to crete a numpy like libary that uses all cores and works on datasets not fully in memoery
- dask doe sthis in ga genreal way

blocking algorithms

- blocked algorithms compute a large result with may small computations,
- dask is built on blocked algorirms that is breaking operations into many small chunks
- can execute all parts of a graph with the .compute method

array metadata

- dask array objects have the following information
    1. a dask graph
    2. infomration aout shape and chunk shape
    3. a name
    4. a data tpye
- a dask array needs to know the size of all internal chunks which can be ragged
- can slice dask arrays by chunks

capatbilties and limitations

- dask works on most numpy functions, but does not work with functions whose output shape can not be determined head of time like where statments

dynamic task scheduling

- dask has dynamic task scheduling
- more or less when a worker completes a task, the runtime state is updated, and a new task is chosed from among the set of ready to run tasks
- dask does not push intermeidate results to disk like map reduce does
- they use lifo task execution
- user can also modify the scheudler for there task if they so desire
- there are also numerous scheudlers including single machine, multi thread multi process and distributed
- then they talk about some benchmarking

other collections

- the dask libary containes parallel collections other than dask.array including dask.bag and dask.dataframe

bag

- a bag is an unorderd colleciton with repeates.
- is really good for initial data cleaning because it uses proefmance form other well estbalished tools and adds pararllism

data frame

- the dask data frame module implements a large datafrmame out of many pandas datafames.
- the interface is bassed off of pandas
- daa frames are powerfull but can not achive the same parallel preformance as arrays or bags

- data frames can efficently do computation on partioned data sets as well as allong axis if they are in the same block
- dask data frames are helpfull as they let users easily acess datasets that are larger than memory using the pandas interface

dask for general computation

- daks can work well with sopisticaed parallel algorirhsm on multi core machines
-