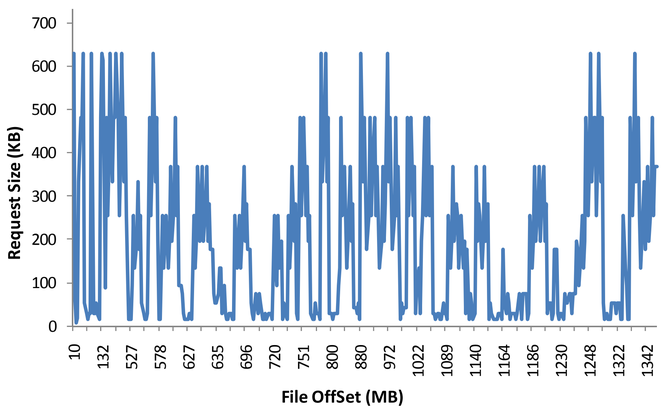
Trace-based Adaptive Data Layout Optimization for Parallel File Systems

Huaiming Song[1] Xian-He Sun[1] Hui Jin[1] Yong Chen[2]
[1] Illinois Institute of Technology [2] Oak Ridge National Laboratory

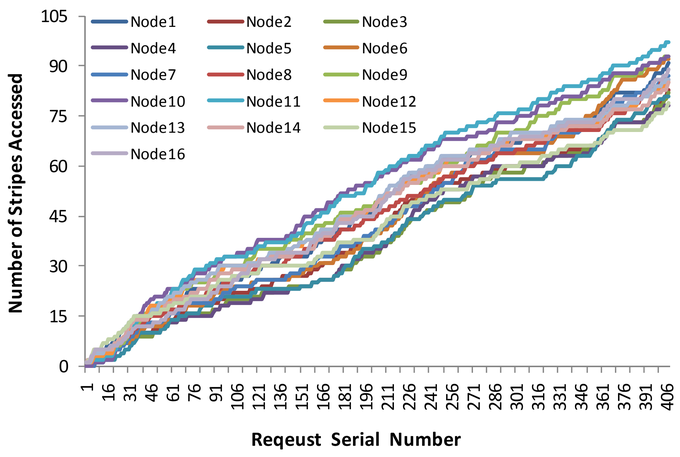
Non-uniform Data Access Problem

- Data size grows rapidly
  - The whole file system: peta-scale/exa-scale
  - A single file: several terabytes or even beyond
- Non-uniform data access
  - Different access patterns for different parts of data files
    - Request size could be large or small
    - The number of concurrent I/O processes might be varied
  - Unified strip size for whole file cannot achieve optimal I/O bandwidth for all requests
  - Data access might be not balanced for all I/O servers

Case Study: Chombo trace



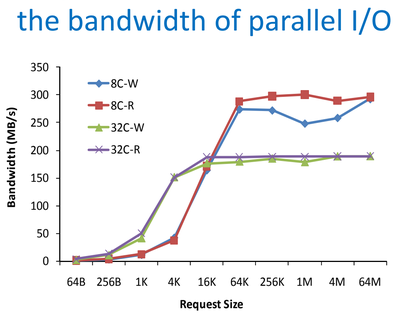
Request sizes at different file offsets of Chombo trace. We can see that request size changes a lot on different parts of the file.



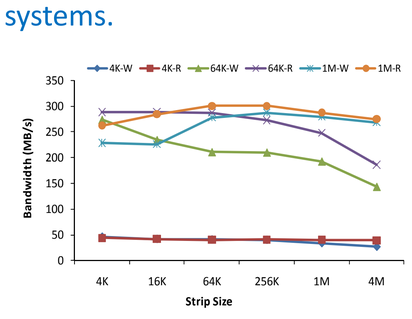
The number of strips accessed on different I/O servers. Assume that file is striped in a 16-node parallel file system with 64KB strip size (default value of PVFS2). We can see that data access is not balanced for all I/O servers.

Two Key Factors Affect I/O Performance

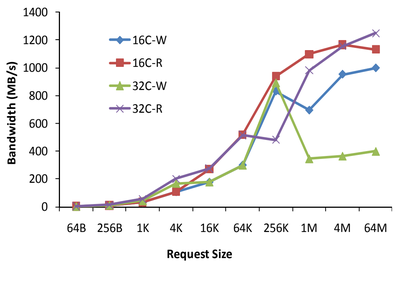
Request size and **strip size** are two key factors that affect the bandwidth of parallel I/O systems.



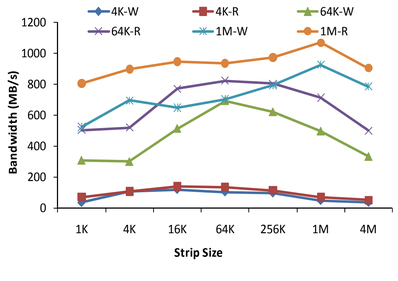
(a) Ethernet: fixed strip size



(b) Ethernet: fixed request size



(c) InfiniBand: fixed strip size

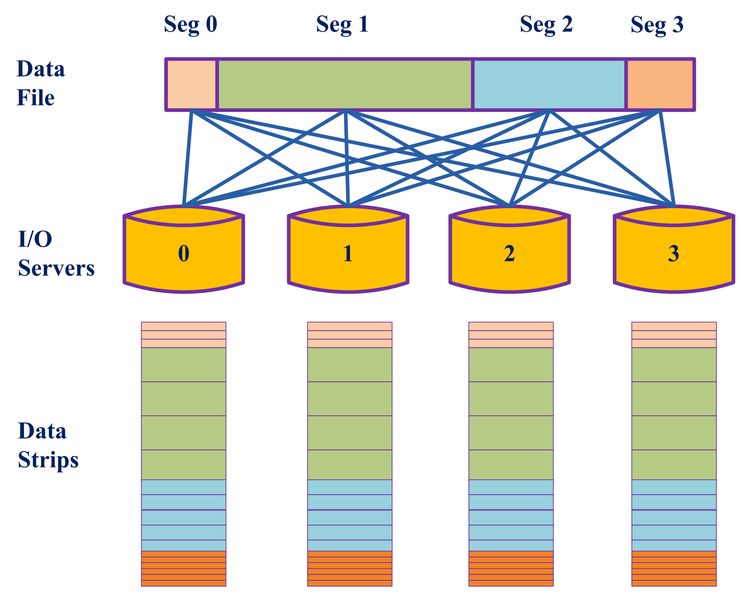


(d) InfiniBand: fixed request size

Fine-grained Layout Optimization

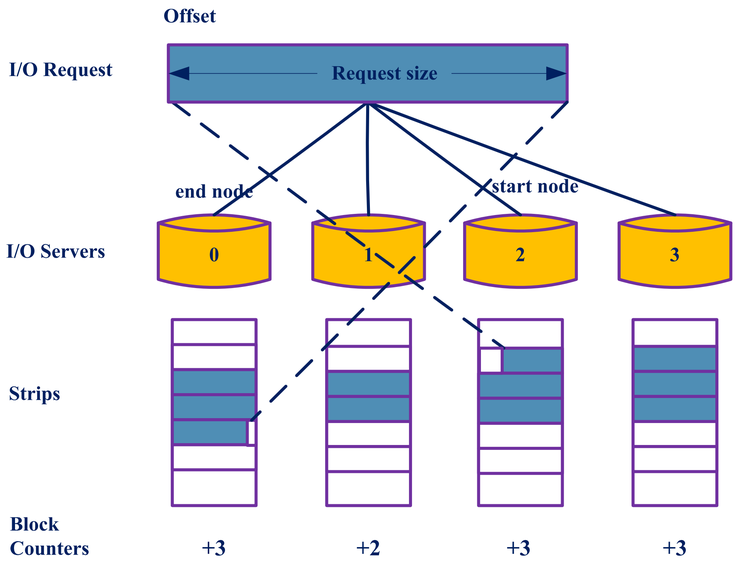
- Idea of fine-grained layout optimization
  - Large file is divided into small segments
  - Individual strip size for each segment to achieve higher I/O bandwidth
  - Take data access balance into consideration
- Basic approach (5 steps)
  1) divide large file into small segments by a fixed block size: e.g. 64MB or 128MB;
  2) calculate the average access pattern for each segment based on data access traces;
  3) analyze data access cost for each segment respectively according to parallel I/O cost model, calculate the optimal strip size for each segment;
  4) count the number of strips accessed on all I/O servers for each segment. Choose a proper strip size close to optimal strip size (in step 3) if not balanced;
  5) combine adjacent segments into a larger segment if the two have the same or very close strip size.

Data Layout Description



Data layout metadata is described as an <offset, strip-size> pair list, e.g. <0, 16KB: 4MB, 1MB: 1024MB,64KB: 1280MB, 4KB>.

Data Access Load Balance

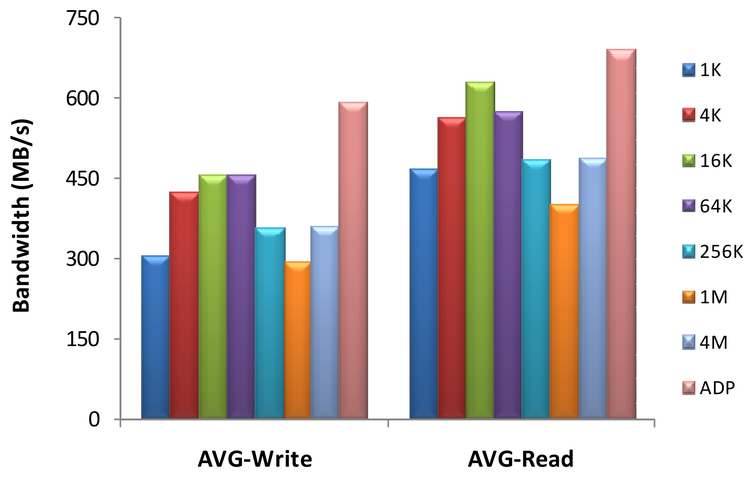


- Data access workload measurement: the number of strips accessed by all requests
  - Assign a counter for each I/O server
  - Update the values of counters for each request
- Workload imbalance

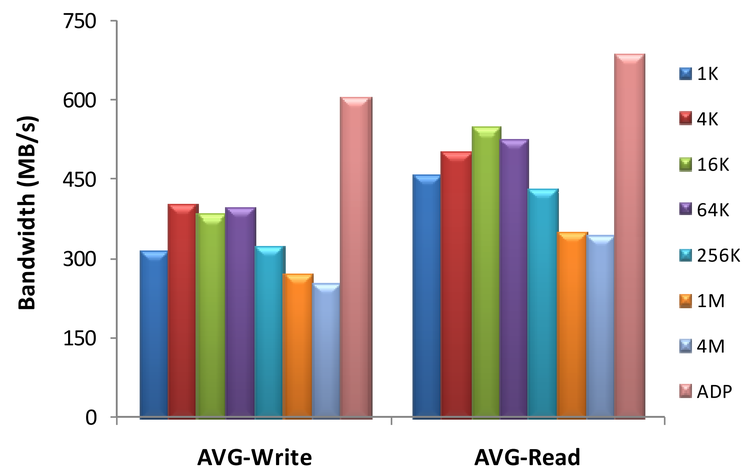
$$\sigma = \frac{L_{max}}{L_{avg}} - 1$$

- Different strip sizes have different effects to workload balance on all I/O servers

Preliminary Experimental Results



Average bandwidth of mixed IOR workloads (workloads balanced). Writing performance of the proposed layout strategy can achieve about 29% to 97% improvement, while reading performance can achieve 10% to 71% improvement.



Average bandwidth of mixed IOR workloads (workload imbalance=0.2 when strip size is 64KB). The trace-based adaptive layout strategy can achieve 51% to 123% improvement with writing performance and 25% to 93% improvement with reading performance.

Conclusions and Future Work

- Conclusions
  - Propose a fine-grained segment-level data layout strategy for applications with non-uniform data access patterns
  - Present the basic approaches of the proposed fine-grained layout schema
  - The new approach also takes data access balance into consideration
  - Preliminary experimental results have verified the effectiveness of our idea
- Future Work
  - We plan to refine data access cost model, for estimating the optimal strip size of each file segment more precisely
  - We plan to work out fined-grained data layout schemas for some specific applications

The authors are thankful to Yanlong Yin of Illinois Institute of Technology, Dr. Rajeev Thakur, Dr. Robert Ross and Samuel Lang of Argonne National Laboratory for their constructive and thoughtful suggestions toward this study. This research was supported in part by National Science Foundation under NSF grant CCF-0621435 and CCF-0937877.

Contact Information: Huaiming Song(hsong20@iit.edu), Xian-He Sun(sun@iit.edu), Hui Jin(hjin6@iit.edu), Yong Chen(cheny@ornl.gov)