

Linea, A New Assembly System

Mon-Chaio Lo (1), Gary Montry (2), Christie Robertson (1), Todd Smith (1)

1) Geospiza, Inc., Seattle, WA <http://www.geospiza.com>

2) Southwest Parallel Software, Albuquerque, NM <http://www.spsoft.com>

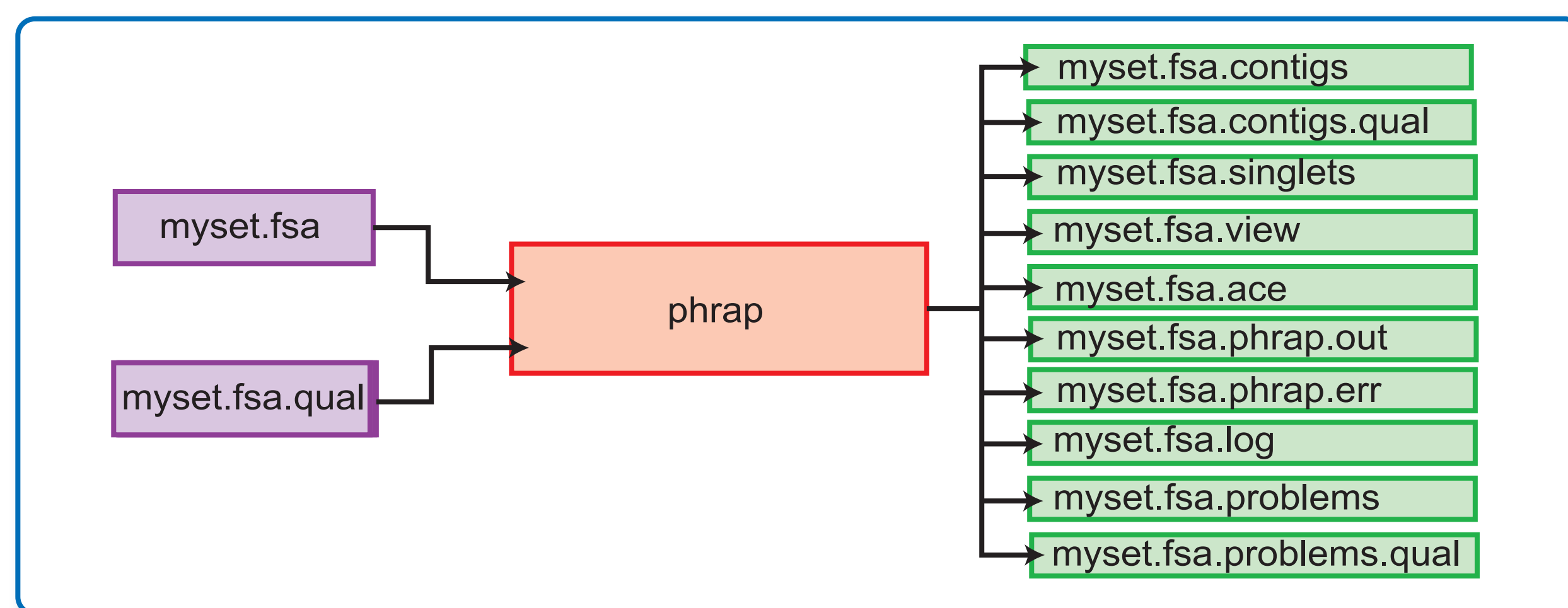


Southwest Parallel Software
Providing High-Performance Software Solutions



Assembly Systems - The Old Way

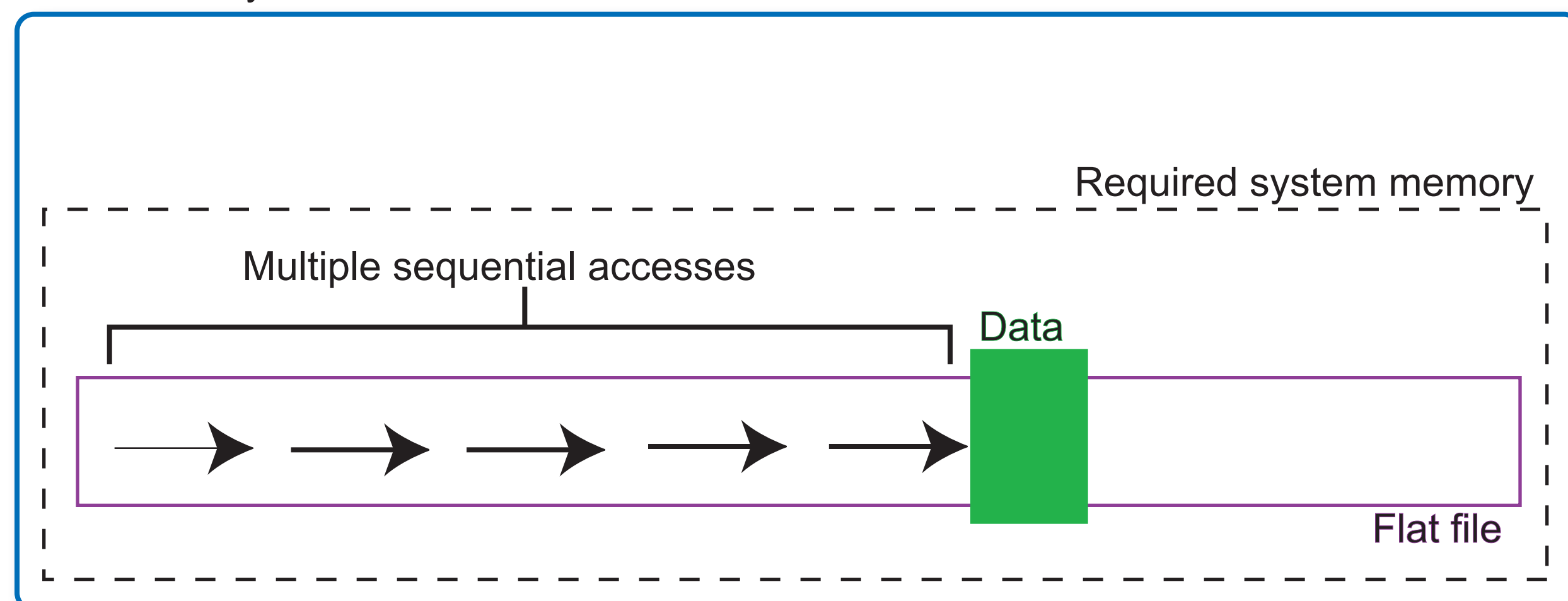
The traditional assembly paradigm organizes information into a number of flat files, each with their own format. Notice that the input and output file formats differ, meaning that phrap output cannot be used as phrap input.



The Problem with Flat Files

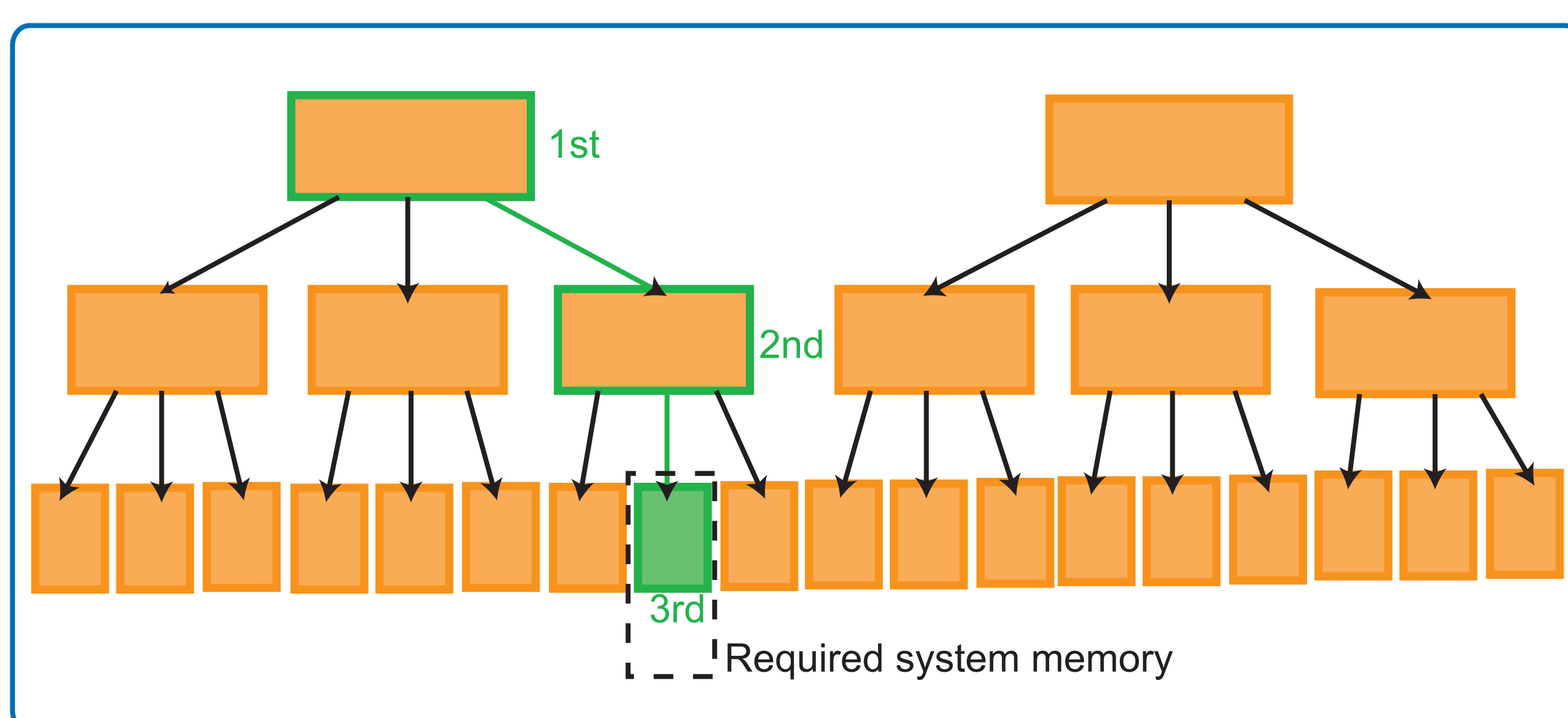
Flat files only allow sequential access, which means:

- 1) File access is slow, especially searching for specific elements inside a file
- 2) Assembly memory requirements are higher because the entire file must be kept in memory

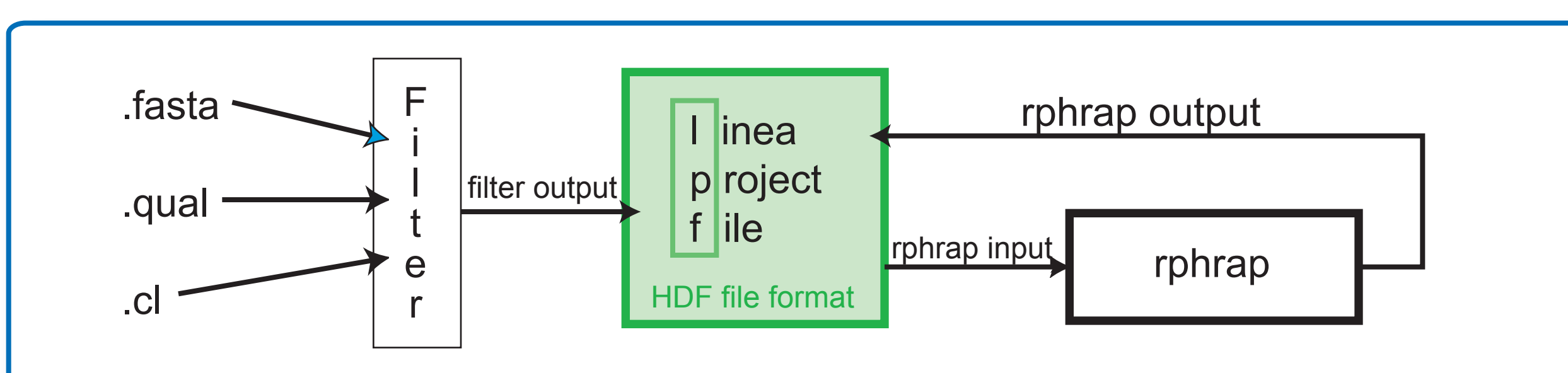


Solution: Binary file format (HDF5)

Sorted B-Tree allows any data to be extracted with a maximum of three disk accesses



The new binary file format allows for the development of an integrated storage solution for next-generation bioinformatics software, such as rphrap, Geospiza's new phrap-like assembler

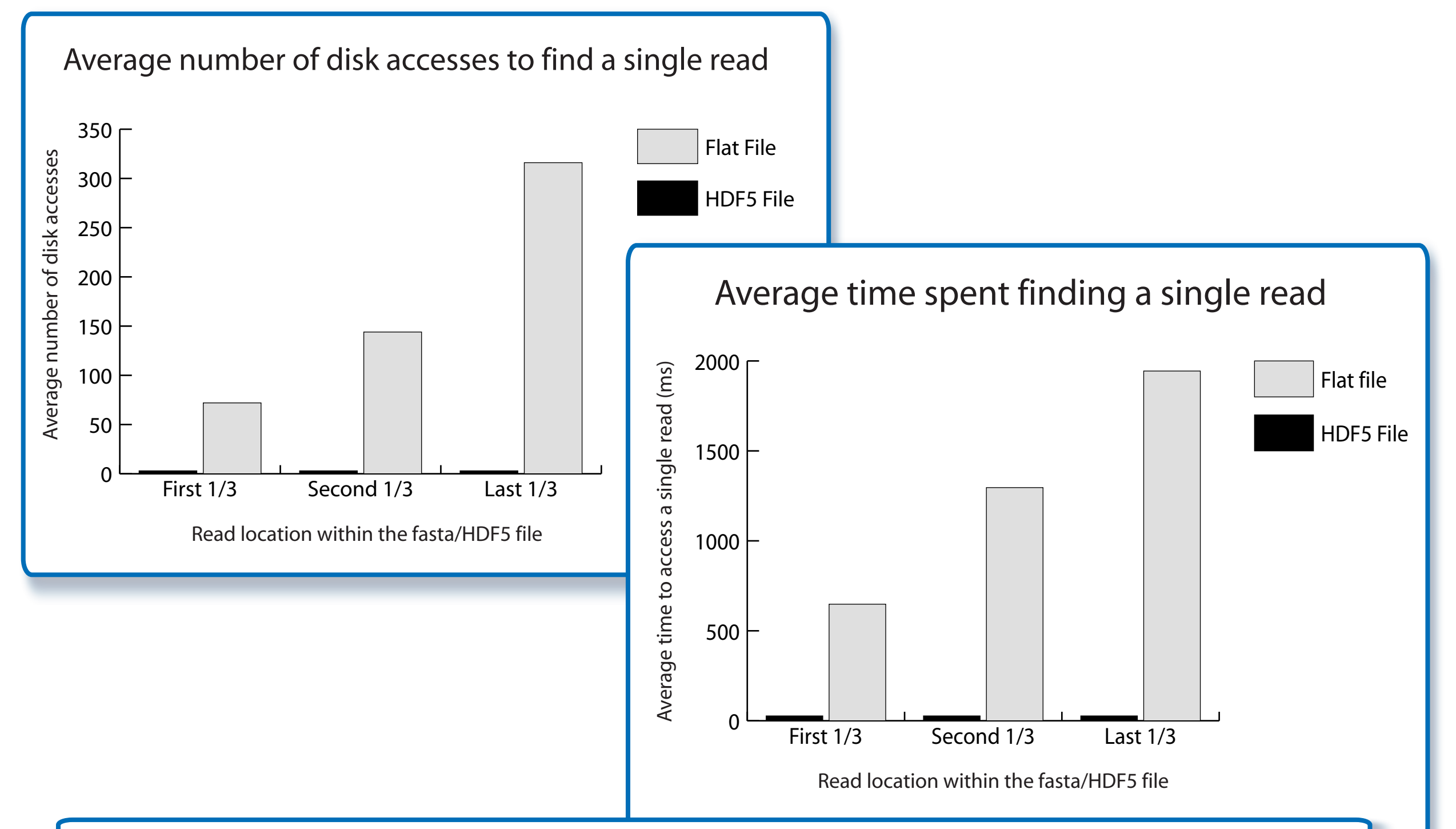


Other HDF5 benefits

- Native support of matrix data
- Open source

Performance Comparison

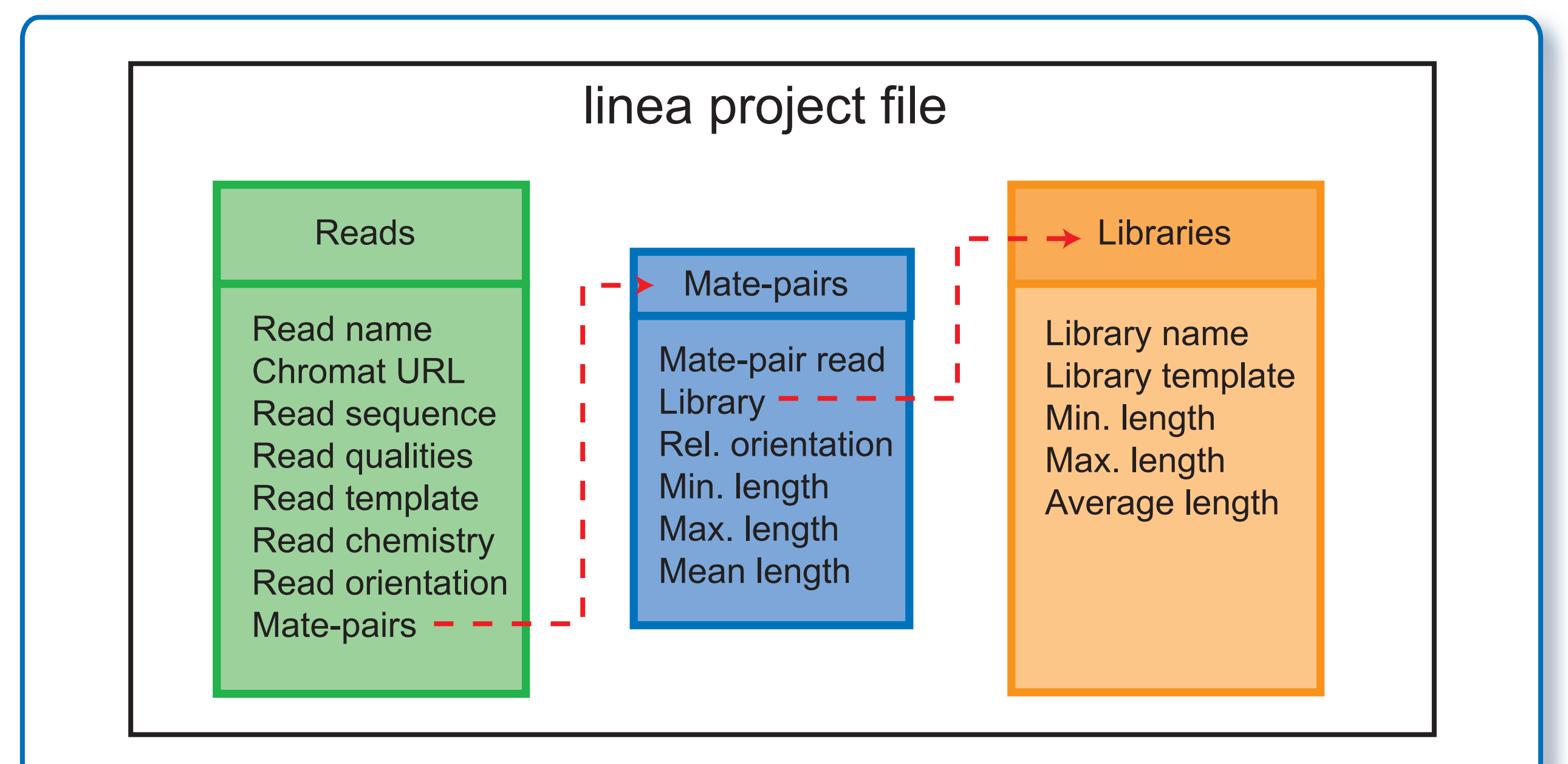
Graphs assume a typical fasta file containing 5000 reads of 700 bp each. Disk parameters are set at 8129 bytes/block with an average access time of 9ms.



Average performance increase of 4800%!

Linea - The New Way

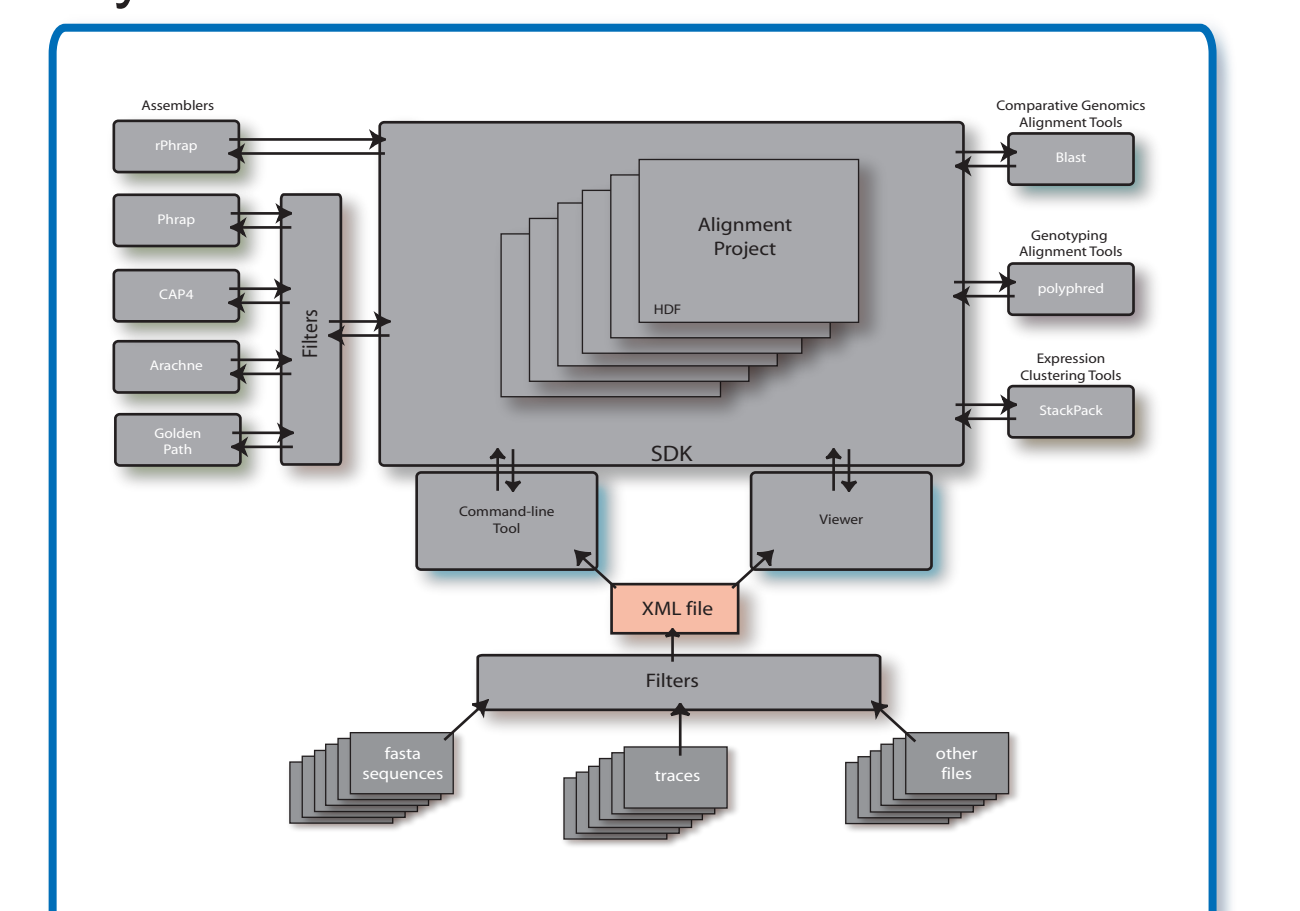
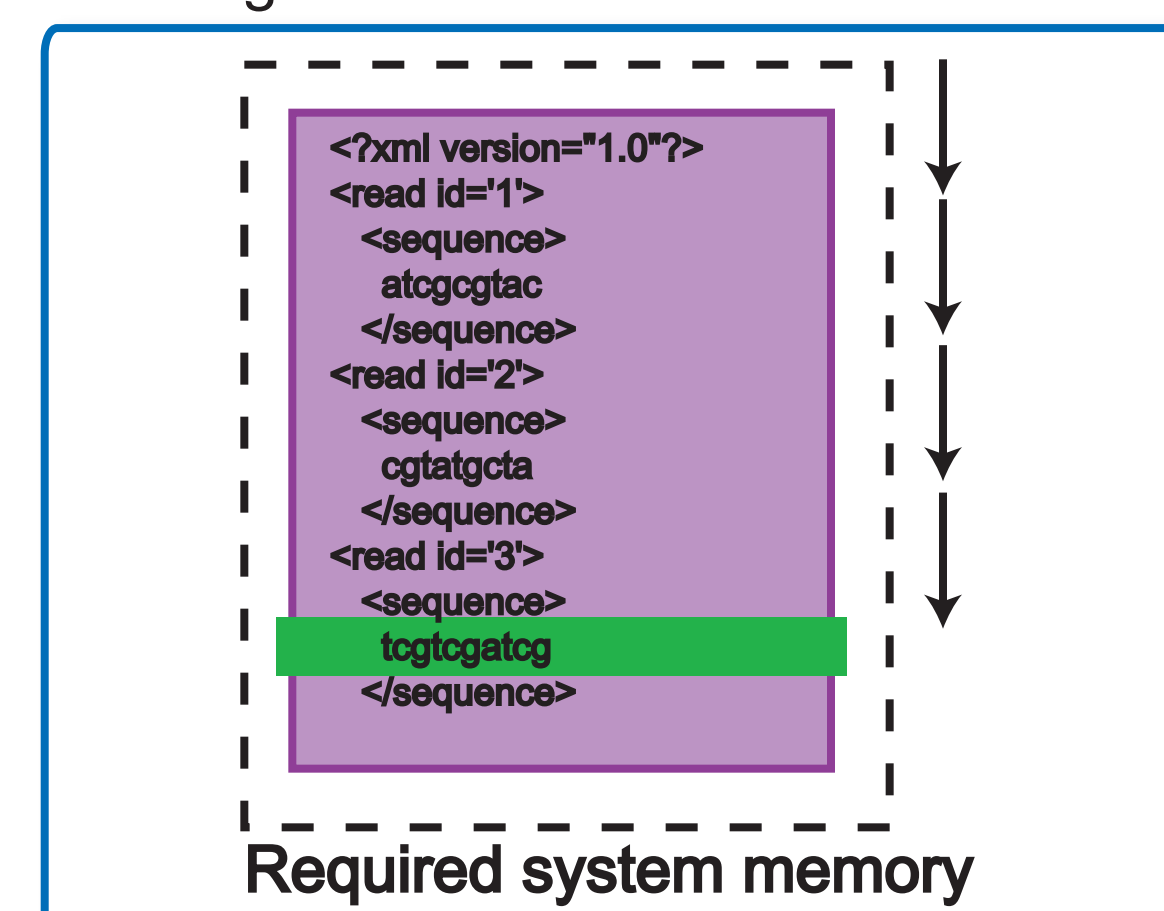
The linea assembly system is a new assembly system that addresses these problems by eliminating the use of flat files for data storage. At its core is the linea project file, a structured, extensible binary file format based on HDF5.



Aside: Correct XML Usage

An XML file is a flat-file, and so has all the deficiencies mentioned previously. This makes it a poor candidate to use as a storage format.

linea uses XML where it really shines, as a transport layer language. XML works well here because data in the transport layer does not need to be random access.



HDF5 - The Next Generation of HDF:
<http://hdf.ncsa.uiuc.edu/HDF5/>
December-30-2003 (Date of Access).

We are grateful to the National Institutes of Health for SBIR grant #R44 HG02244-02, which supports this project.

Geospiza, Inc. and the finch logo are registered trademarks of Geospiza incorporated. All other trademarks, service marks and registered trademarks appearing herein are the properties of their respective owners and are hereby acknowledged.
Presented at Plant and Animal Genomes XII Conference, San Diego CA, 2004