Gfarm Grid File System

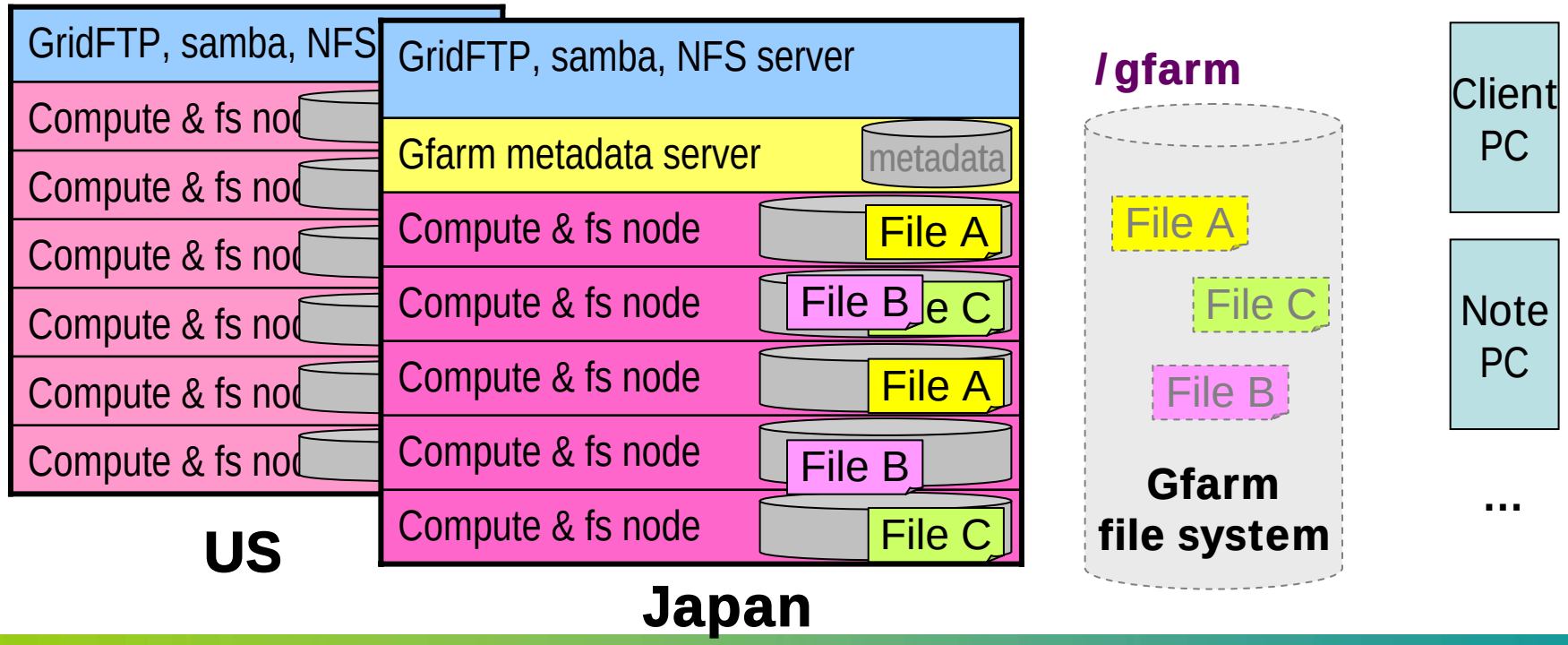
Osamu Tatebe
University of Tsukuba

Gfarm Grid File System

- **Next-generation network file system**
  - It is a distributed file system that federates a local disk of cluster nodes or Grid nodes
  - It can be shared among all cluster nodes and clients
    - Just mount it as if it were high-performance NFS
  - It provides scalable I/O performance wrt the number of parallel processes and users
  - It supports fault tolerance and avoids access concentration by automatic replica selection
  - Open Source Software

Gfarm Grid File System (2)

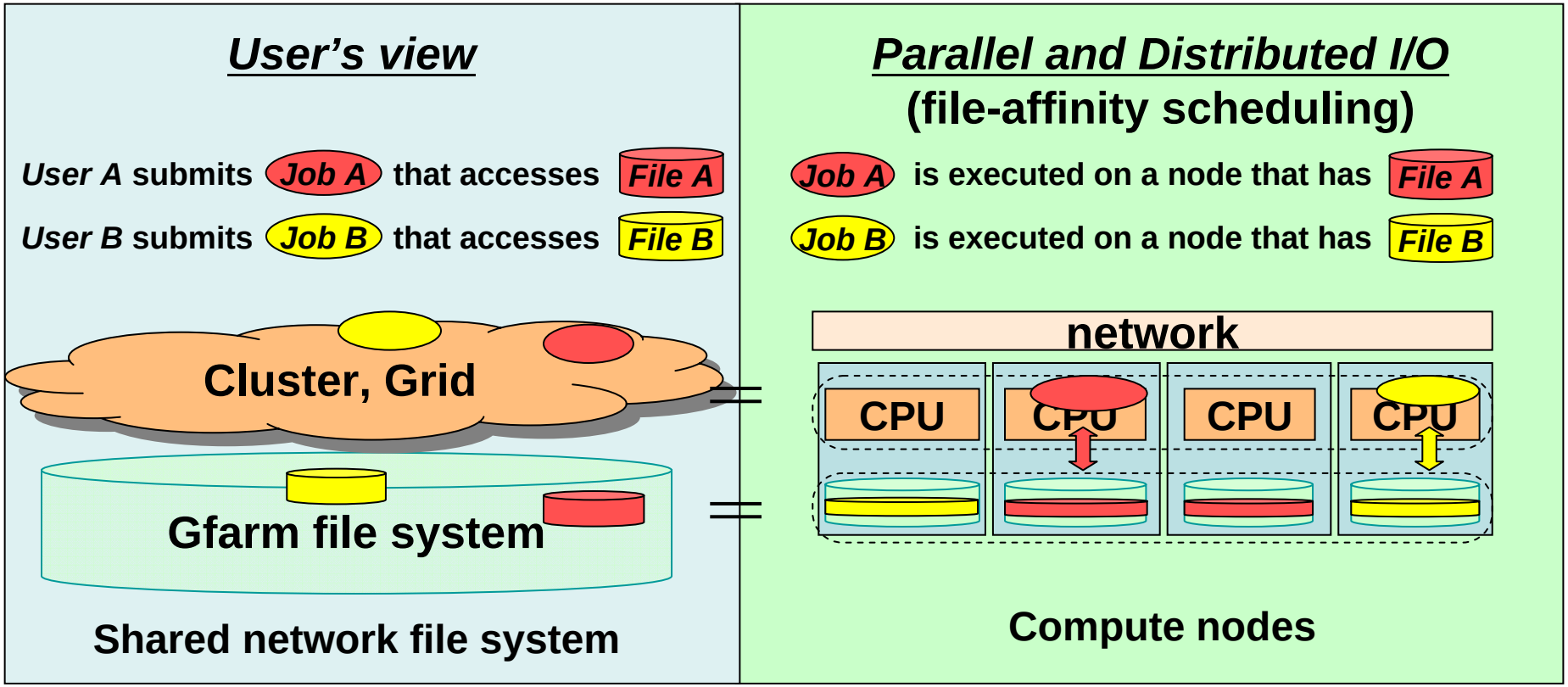
- Files can be shared among all nodes and clients
- Physically, it may be replicated and stored on any file system node
- Applications can access it regardless of its location
- File system nodes can be distributed



Scalable I/O Performance

- **Decentralization of disk access putting priority to local disk**
  - When a new file is created,
    - Local disk is selected when there is enough space
    - Otherwise, near and the least busy node is selected
  - When a file is accessed,
    - Local disk is selected if it has one of the file replicas
    - Otherwise, near and the least busy node having one of file replicas is selected
- **File affinity scheduling**
  - Schedule a process on a node having the specified file
    - Improve the opportunity to access local disk

Scalable I/O Performance



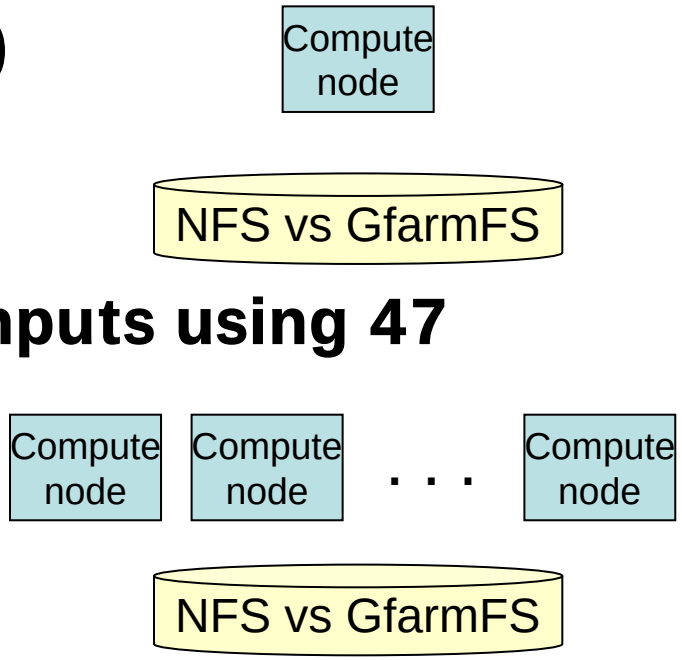
Do not separate storage and CPU (SAN not necessary)

Move and execute program instead of moving large-scale data

Scalable file I/O by exploiting local I/O

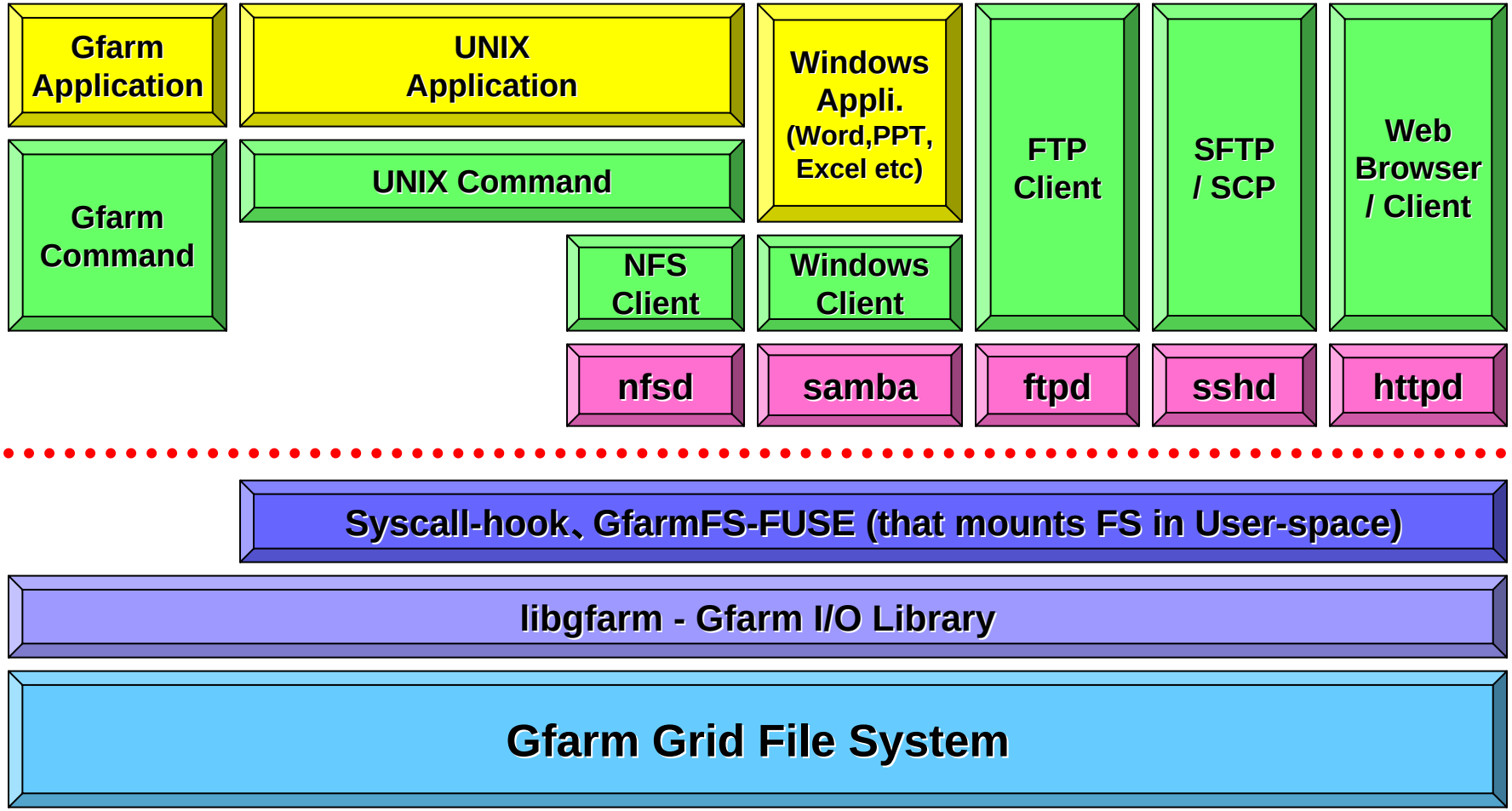
Example - Gaussian 03 in Gfarm

- **Ab initio quantum chemistry Package**
  - Install once and run everywhere
  - No modification required to access Gfarm
- **Test415 (IO intensive test input)**
  - 1h 54min 33sec (NFS)
  - 1h 0min 51sec (Gfarm)
- **Parallel analysis of all 666 test inputs using 47 nodes**
  - Write error! (NFS)
    - Due to heavy IO load
  - 16h 6m 58s (Gfarm)
    - Quite good scalability of IO performance
    - Longest test input takes 15h 42m 10s (test 574)



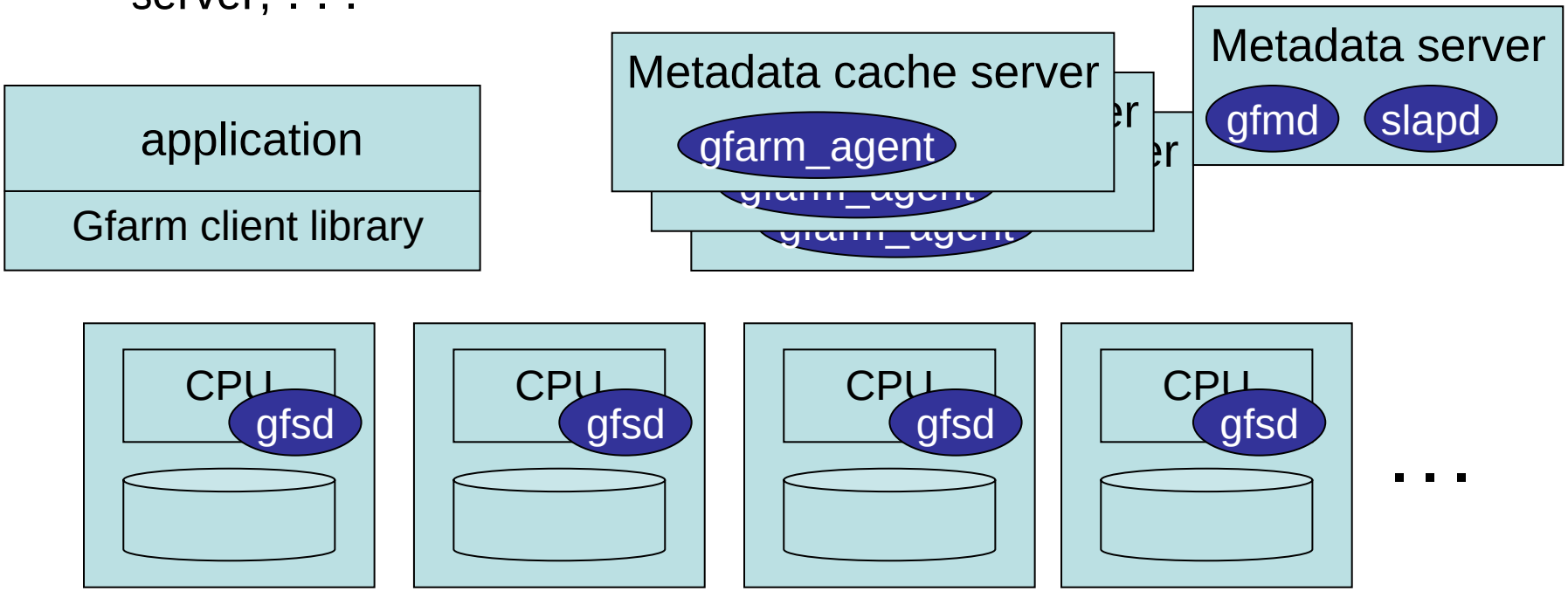
*Gfarm consists of local disks of compute nodes

Software Stack



Gfarm™ File System Software

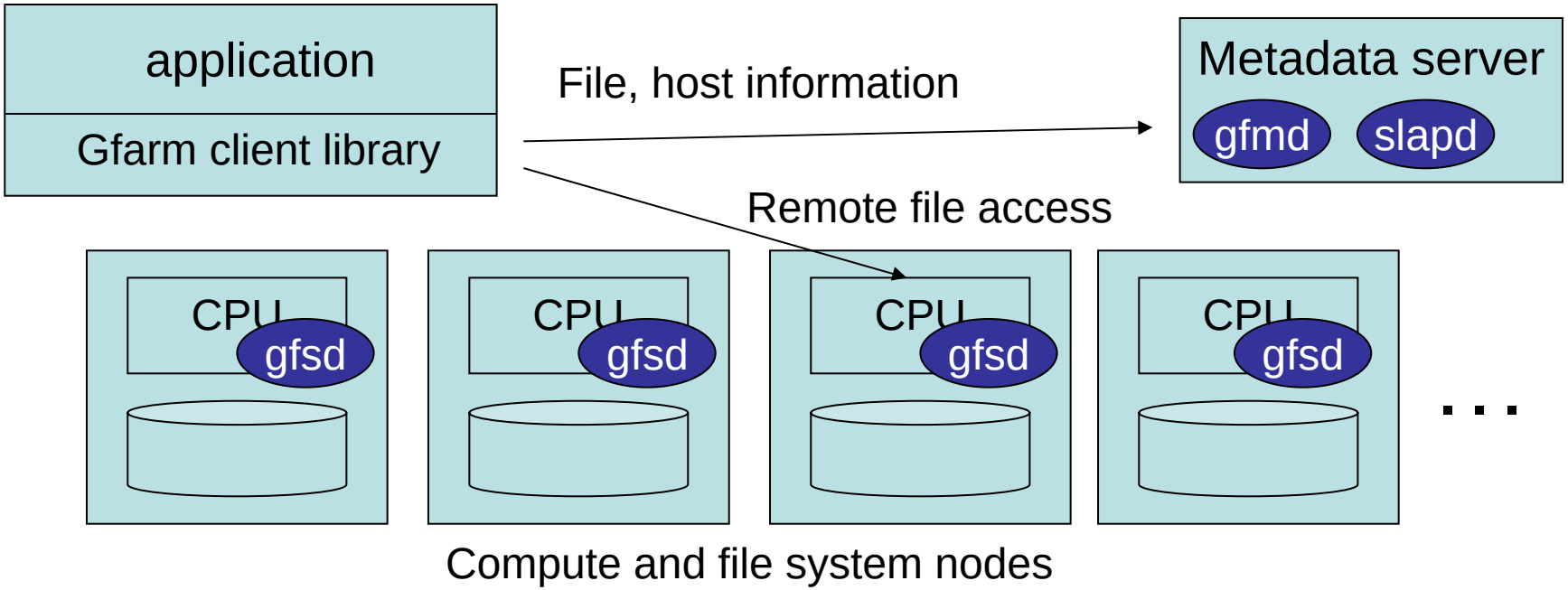
- **Open source development**
  - Gfarm™ version 1.3.1 released on Aug 7th, 2006 (http://datafarm.apgrid.org/)
  - A shared file system in a cluster or a grid
  - Accessibility from legacy applications without any modification
  - Standard protocol support by scp, GridFTP server, samba server, . . .



Compute and file system nodes

Gfarm™ File System Software (2)

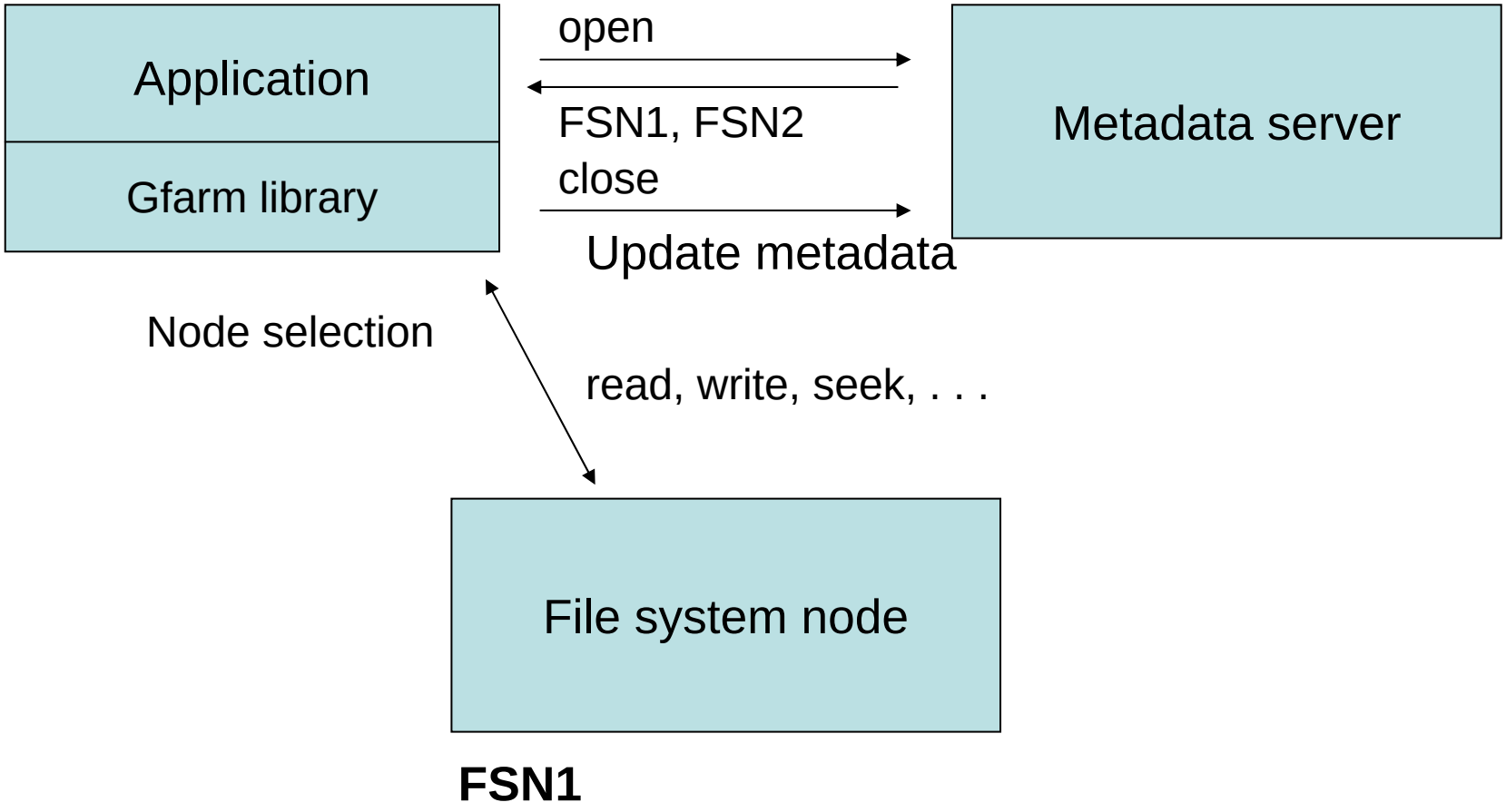
- **libgfarm – Gfarm client library**
  - Gfarm API
- **gfmd, slapd – Metadata server**
  - Namespace, replica catalog, host information, process information
- **gfsd – I/O server**
  - Remote file access



Demonstration

- **File manipulation**
  - cd, ls, cp, mv, cat, . . .
  - grep
- **Gfarm command**
  - File replica creation, node & process information
- **Remote (parallel) program execution**
  - gfrun prog args . . .
  - gfrun -N #procs prog args . . .
  - gfrun -G filename prog args . . .

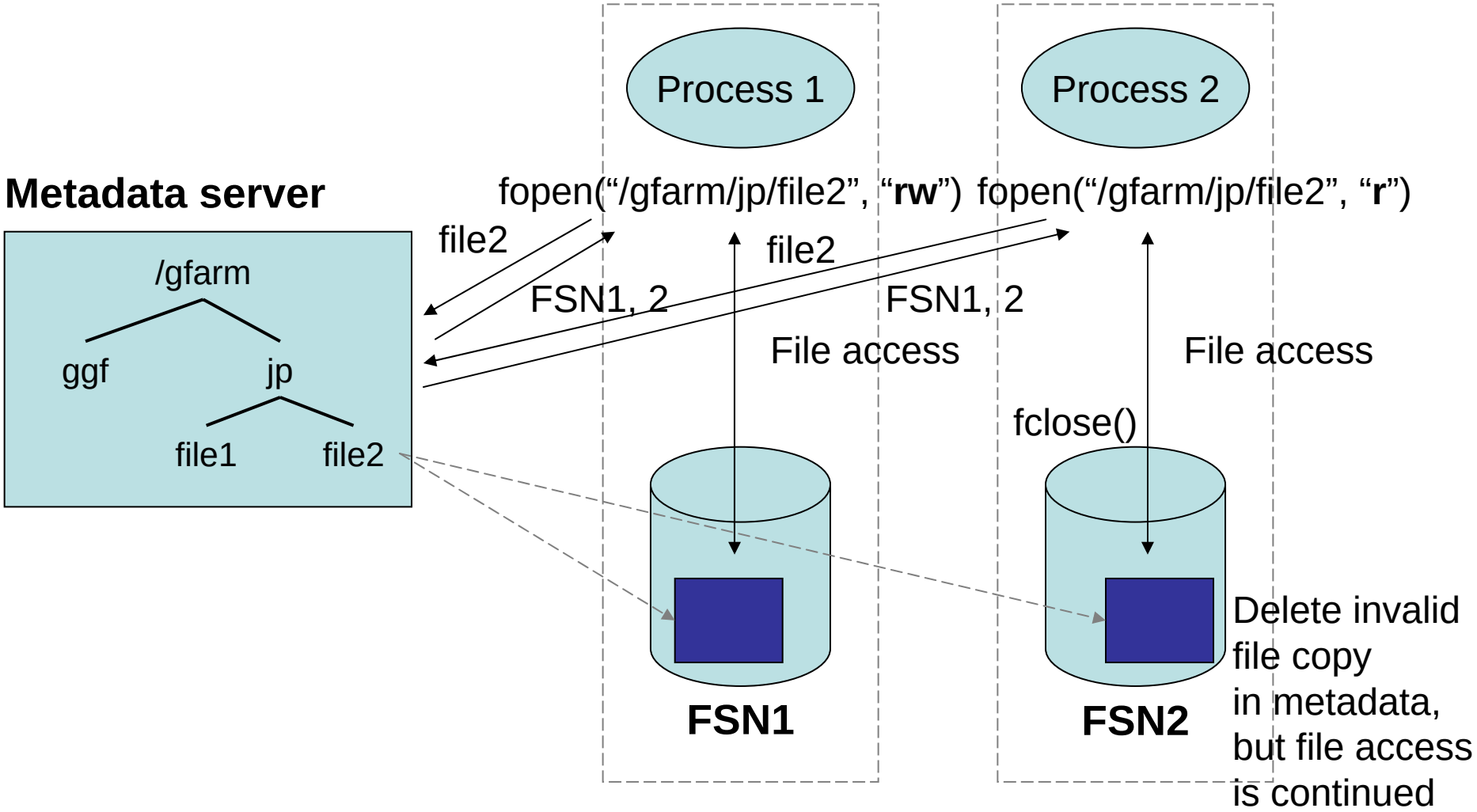
I/O sequence example in Gfarm v1



Opening files in read-write mode (1)

- **Semantics in Gfarm version 1**
  - Updated content is available only when opening the file after a writing process opens the file
  - No file locking (will be introduced by version 2)
    - It supports exclusive file creation (O_EXCL)

Opening files in read-write mode (2)



Access from legacy applications

- **Mounting Gfarm file system**
  - GfarmFS-FUSE for Linux
  - Need volunteers for other operationg systems
- **libgfs_hook.so – system call hooking library**
  - It hooks open(2), read(2), write(2), ...
  - When it accesses under /gfarm, call appropriate Gfarm I/O API
  - Otherwise, call ordinal system call
  - Re-link not necessary by specifying LD_PRELOAD
  - Higher portability than developing kernel module

Advanced Feature

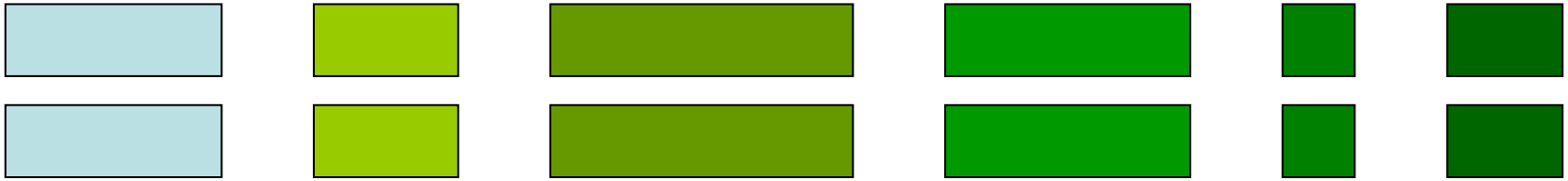
- **Parallel processing in distributed environment**
  - File that consists of multiple file fragments
  - File view
  - File affinity scheduling & local file view

Files in Gfarm File System

- **A file in Gfarm file system may consist of one or more file fragments**
  - Divide a large-scale file into multiple fragments
  - Group multiple files

- **Each file fragment can be stored on any file system node**

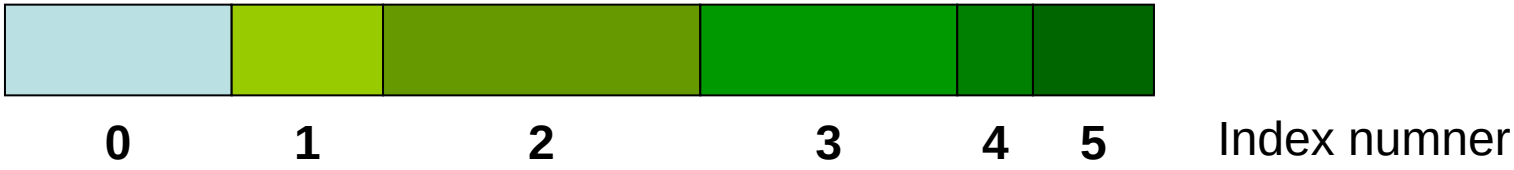
- **Each file fragment can be replicated and stored on any file system node**



File view (1) – Global and index file view

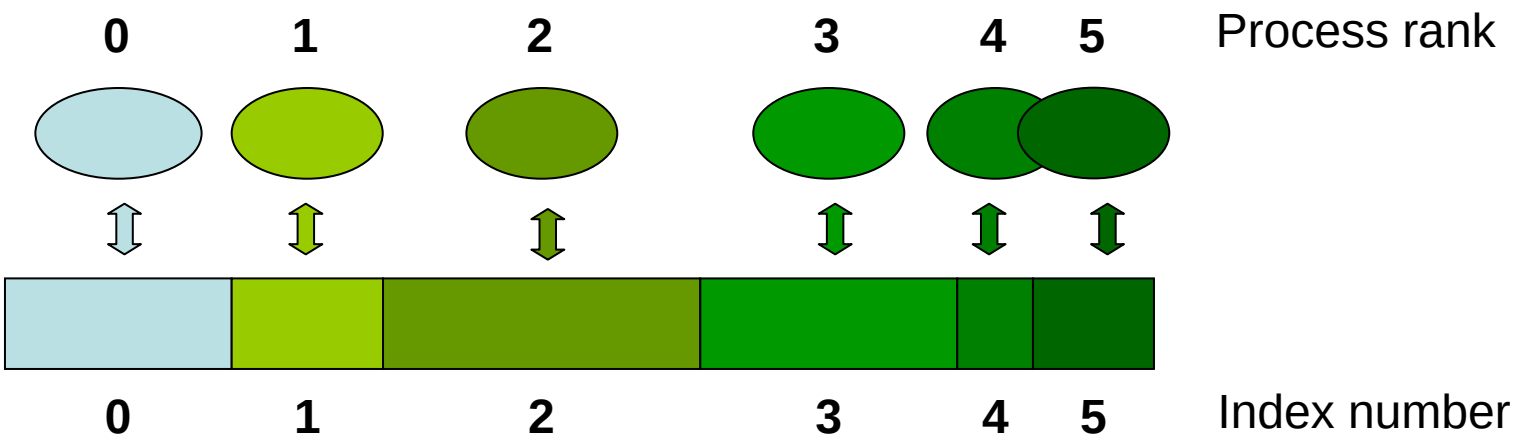
- **Global file view (default)**
  - File is transparently accessed regardless of file fragments

- **Index file view**
  - Only specified file fragment is accessed



File view (2) – Local file view

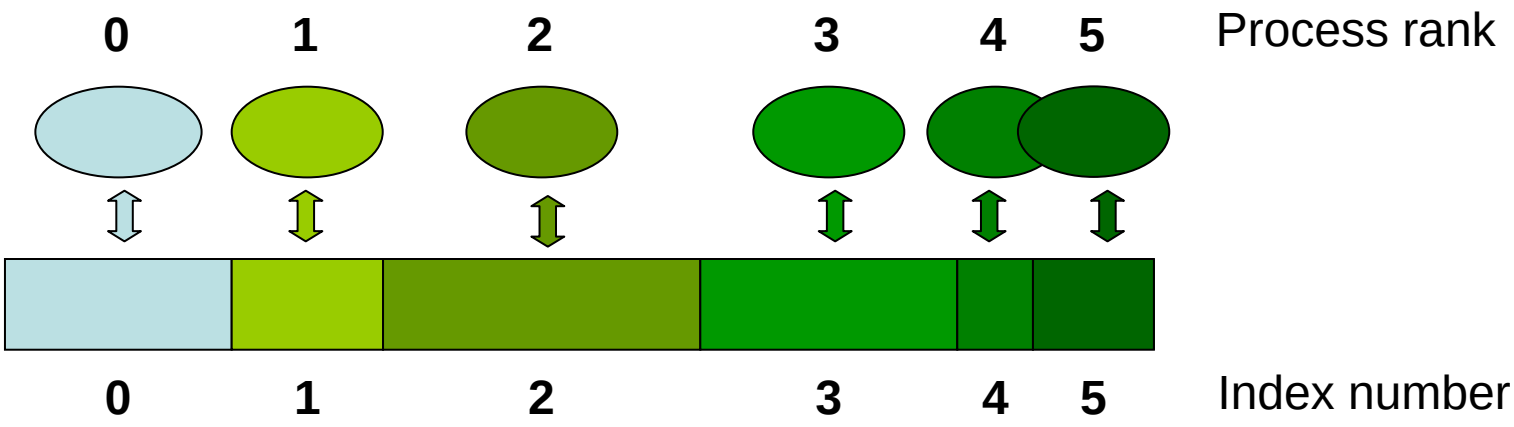
- **Local file view**
  - Special case of index file view
  - Each parallel process accesses the corresponding file fragment



  - Each parallel process accesses independent file fragment
    - A key for scalable parallel I/O performance

File-affinity Scheduling

- **Process scheduling based on file location**
  - Execute the same number of parallel processes as the file fragments on a node having one of the corresponding file fragment copy
  - local file view is a default



% Sequential process is executed when the number of file fragment is one.

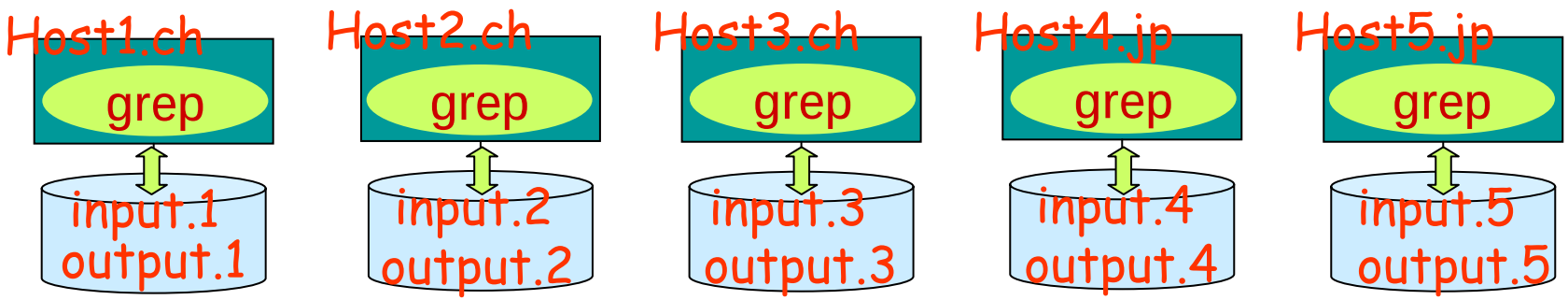
Example of parallel & distributed computing

- **Store a text file to a file having five file fragments**

  % gfimport_text –N 5 –o input input

- **Execute parallel grep**

  % gfrun –G input grep *regexp* input

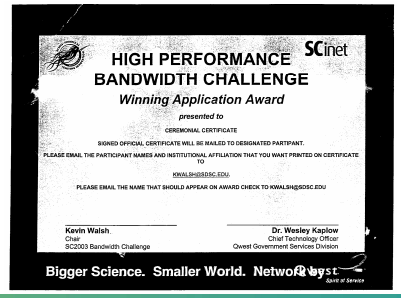


File-affinity scheduling

More Feature of Gfarm Grid File System

- **Commodity PC based scalable architecture**
  - Add commodity PCs to **increase storage capacity** in operation much more than **petabyte scale**
    - Even PCs at distant locations via internet
- **Adaptive replica creation and consistent management**
  - Create multiple file replicas to **increase performance and reliability**
  - Create file replicas at distant locations for **disaster recovery**
- **Open Source Software**
  - Linux binary packages, ports for *BSD, . . .
    - It is included in Naregi, Knoppix HTC edition, and Rocks cluster distribution
  - **Existing applications** can access w/o any modification
    - Network mount, Samba, HTTP, FTP, ..

http://datafarm.apgrid.org/



SC03 Bandwidth SC05 StorCloud