Administration and Programming for the Red Storm IO System

May 19 2005

Lee Ward
lee@sandia.gov

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

National Nuclear Security Administration

Sandia National Laboratories

Topics

- Application programming interface
  - POSIX and ASCI Red compatibility
  - New functionality for Red Storm
- Architecture
- Base File System Drivers
- Administration
  - Initialization and Startup
  - Shutdown
  - User extension

Sandia National Laboratories

The xtio.h include file

- Defines types and calls not found in the normal POSIX includes
  - For instance ioid_t and iread, iwrite, ireadx, etc.
- Must reference this in your program source if using any calls other than those found in POSIX

POSIX Compatibility

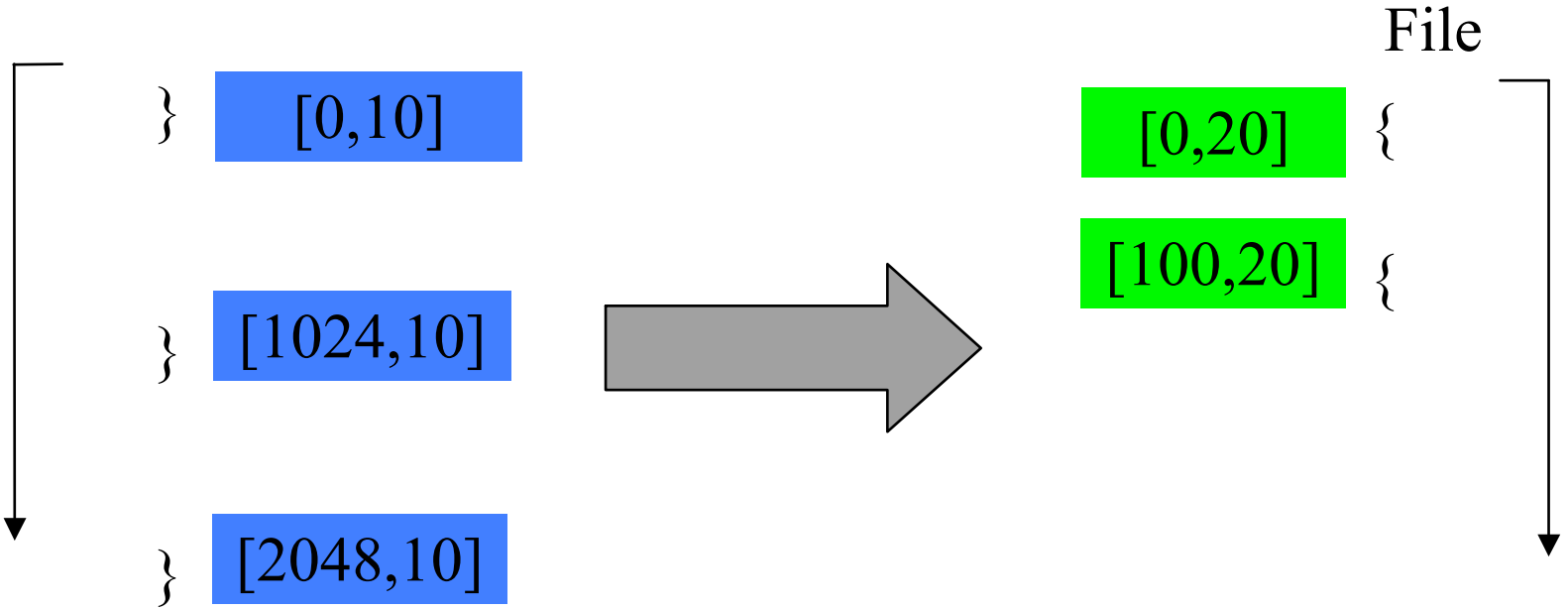
- Most common entry points supported
  - In the most commonly used ways
  - open, close, read, write, mkdir, unlink, etc.
    - Full list found in the paper
- Not true POSIX
  - The calls are defined but are not complete in all cases
    - FS drivers have a lot of influence
- File descriptors are not shared between compute nodes
  - O_APPEND, for instance may not function as expected

ASCI Red Compatibility

- Asynchronous; iodone and iowait from Red
- Async calls return a derived type called ioid_t though
- The read and write call variants are completely supported
  - For instance; read, iread, ireadv, etc.
- Must *not* count on shared file descriptors and pointers

Sandia National Laboratories

Strided IO via ireadx and iwritex



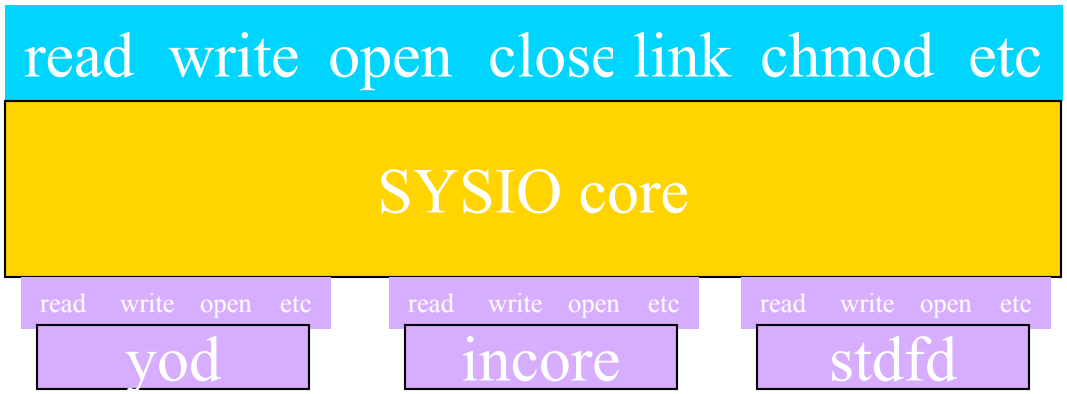
- Gather to scatter in this operation
- The two lists are reconciled
- 30 bytes transferred

Sandia National Laboratories

Other New Functions

- For completeness
  - For instance, new standard introduces pread for instance
- All datapath variants defined
  - 'i' meaning asynchronous
  - 'p' (from POSIX) meaning “position” first
  - 'v' meaning scatter/gather from/to memory
  - 'x' meaning scatter/gather between memory and the file address space
    - Strides

SYSIO architecture



- VFS in user space abstracts common function
  - But namespace maintenance is local
    - So, a global namespace isn't guaranteed
- File system driver register with the core
  - Activated at mount time
  - User extensible

Basic File System Drivers

- yod
  - Basic, SUNMOS original, function shipping interface to launcher IO capabilities
- incore
  - In memory scratch
    - Used for name space assembly
    - Typically root and automount directory templates
- stdfd
  - Hooks driver for 'C' standard input, output, and error descriptors

Initialization

- The sysio_init function initializes and readies the VFS layer
  - Red Storm calls this implicitly from the application run time startup function
- The sysio_boot function is used to enable options and craft name space
  - Typically from a passed environment variable
  - Red Storm calls it implicitly from application run time start routine
  - The user may call it as well to add other, non-standard drivers, enable debugging, etc.

Automounts

- Parent mount must have automounts enabled
  - MOUNT_F_AUTO option is set
- Create, or find, an empty directory
- Add a file called .mount
- Initialize with automount directive
  - <fstype>:<src>[[\t]+<options>]
- Change permissions of the directory so that the SETUID bit is set

Sandia National Laboratories

Startup

- The sysio_boot function interprets a terse command description
  - Can be called more than once
- Supports trace, namespace, cwd
  - The namespace directive supports creat, chmd, mnt, and open
    - The namespace open directive may deposit data as well
      - Automount directive content
- Red Storm used cwd directive to set the initial working directory

Shutdown

- Use sysio_shutdown
  - Provides clean, graceful exit
    - File systems must work without this as applications do crash and hang
  - Called implicitly at application exit
- Cannot restart SYSIO after shutdown

Extending SYSIO to Support Other File Systems

- SYSIO is user-extensible
- Application registers a new file system driver via _sysio_fssw_register
- Application mounts explicitly or accesses automount referring to the new file system

Sandia National Laboratories