Impact of Gfarm, a Wide-area Distributed File System, upon Astronomical Data Analysis and Virtual Observatory

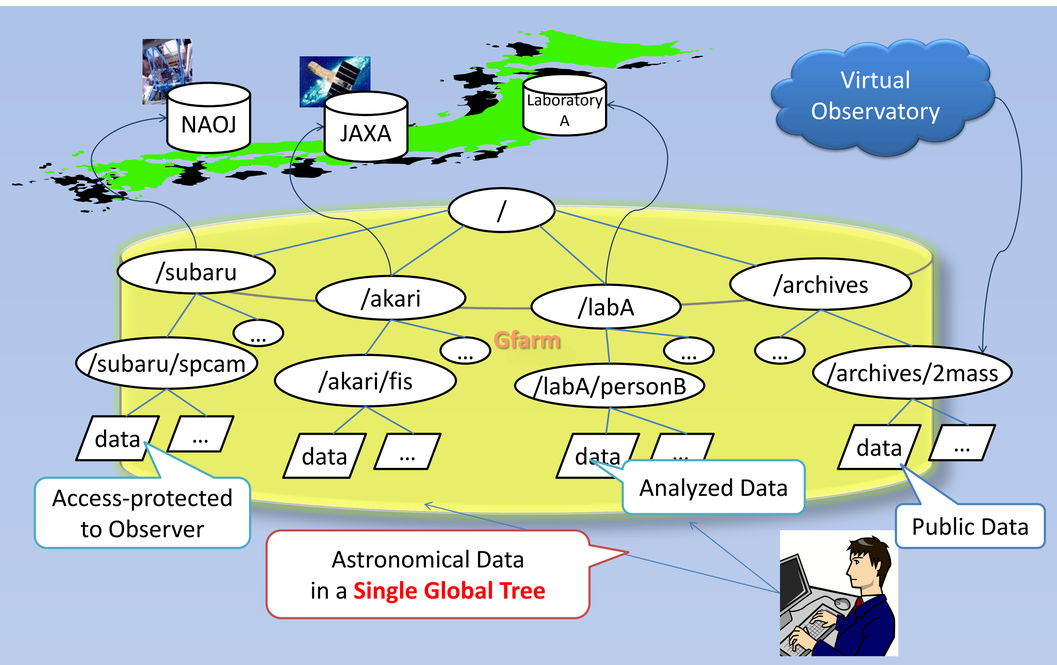
Masahiro Tanaka and Osamu Tatebe (University of Tsukuba)

While 100 TB-scale astronomical data are available through Virtual Observatories, there are still several issues for large-scale data analysis that include transferring a large amount of data and securing enough capacity of storage. We thus propose a VO-capable file system in order to offer easy access to astronomical data, by utilizing Gfarm, a wide-area distributed file system developed as an e-Science infrastructure. Gfarm is a distributed file system that federates storage systems in wide area. It is designed to achieve high reliability and high performance exploiting file replicas and distributed file access. These features facilitate large-scale astronomical data analysis under research collaboration of multiple distant organizations. We discuss file system structure and search method which are compliant with VO standards and the initial performance of data analysis on this system.

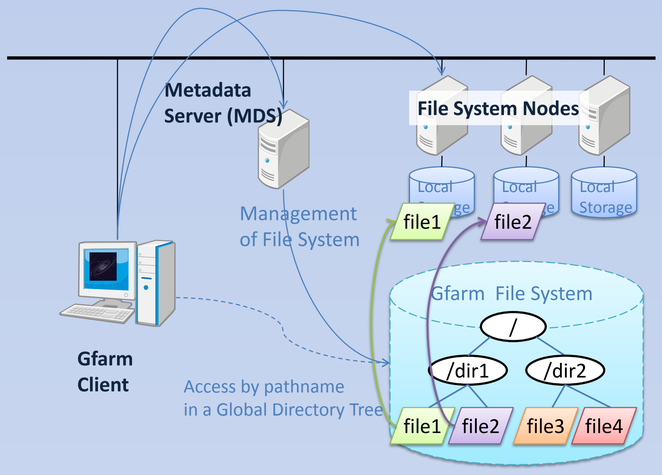
Gfarm

- Wide-area Distributed File System
- Open Source Development
  - http://datafarm.apgrid.org/
- Scalable I/O performance
  - access locality
- Automatic File Replica Selection
  - fault tolerance
  - avoids access concentration
- Fits for Grid Computing

Large area Astronomy data sharing with Gfarm

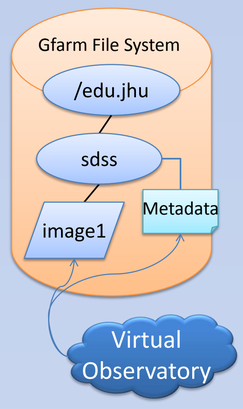


Gfarm Components



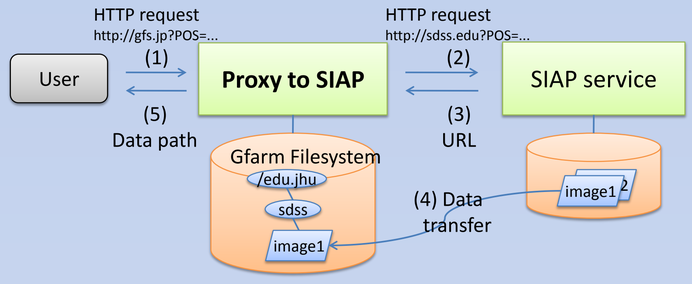
Search into Gfarm with XPath

- Define unique data path
  - use VO ID
    - ivo://edu.jhu/sdss
      → /edu.jhu/sdss
- Data Search
  - VO Metadata (XML)
    - XPath search in Gfarm
- Coordinate Search
  - simple proxy to VO protocol



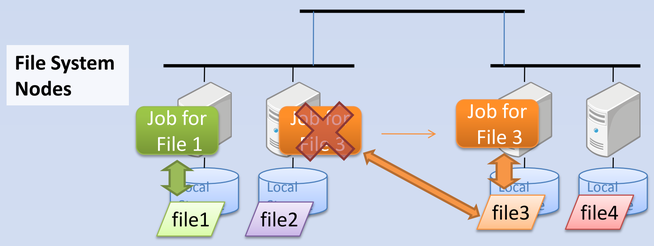
Search into Gfarm with SIAP proxy

- Coordinate search is impossible with XPath.
- Proxy to SIAP service is developed.
  - The Proxy transfers query to original SIAP service.
  - Image data is transferred to Gfarm on demand.



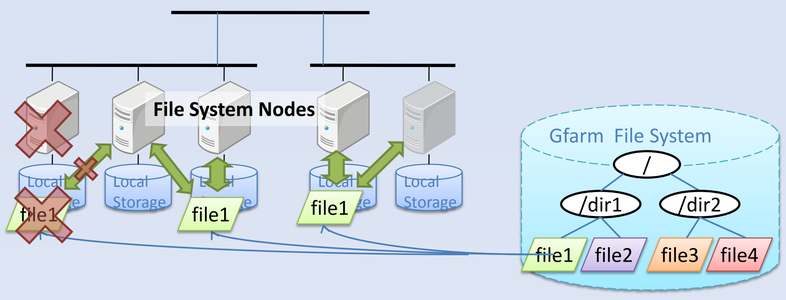
Scalable I/O performance in distributed environment

- Do not separate storage and CPU
- Move and execute program instead of moving large-scale data
- exploiting local I/O is a key for scalable I/O performance



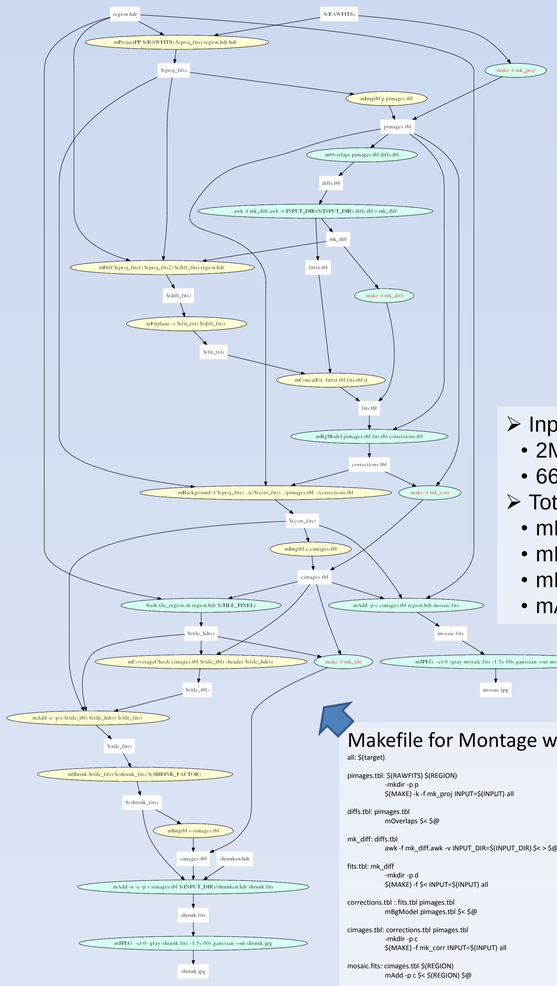
Automatic File Replica Selection

- Files may be replicated and stored in any file system node
  - fault tolerance
  - avoids access concentration



Large-scale data analysis with Gfarm

GXP - parallel grid shell
- workflow is described as Makefile
- parallel job execution
- statistics log - useful for performance evaluation



Makefile for Montage workflow

```
all: ${target}

pimages.tbl: ${RAWFITS} ${REGION}
        -mkdir -p p
        ${MAKE} -k -f mk_proj INPUT=${INPUT} all

diffs.tbl: pimages.tbl
        mOverlaps $< $@

mk_diff: diffs.tbl
        awk -f mk_diff.awk -v INPUT_DIR=${INPUT_DIR} $< > $@

fits.tbl: mk_diff
        -mkdir -p d
        ${MAKE} -f $< INPUT=${INPUT} all

corrections.tbl : fits.tbl pimages.tbl
        mBgModel pimages.tbl $< $@

cimages.tbl: corrections.tbl pimages.tbl
        -mkdir -p c
        ${MAKE} -f mk_corr INPUT=${INPUT} all

mosaic.fits: cimages.tbl ${REGION}
        mAdd -p c $< ${REGION} $@

mosaic.jpg: mosaic.fits
        mJPEG  -ct 0 -gray $< -1.5s 60s gaussian -out $@

t/tile_0_0.hdr: cimages.tbl ${REGION}
        -mkdir -p t
        bash tile_region.sh ${REGION} ${TILE_PIXEL}

simages.tbl: t/tile_0_0.hdr cimages.tbl ${REGION}
        ${MAKE} -f mk_tile INPUT=${INPUT} SHRINK_FACTOR=${SHRINK_FACTOR} all

shrunk.fits: simages.tbl ${INPUT_DIR}/shrunken.hdr
        mAdd -n -e -p s simages.tbl ${INPUT_DIR}/shrunken.hdr shrunk.fits

shrunk.jpg: shrunk.fits
        mJPEG  -ct 0 -gray $< -1.5s 60s gaussian -out $@
```

Performance evaluation of Montage workflow

- Measured using tsukuba cluster of InTrigger Platform
  - Xeon E5410 2.33GHz
  - 2-sockets/node
  - 4-cores/socket
  - 32GB memory
  - 2TB HDD
  - GigEx2 network

Gfarm is 3 times faster
than NFS for 120 parallel procs

- Input file:
  - 2MASS J-band images
  - 6669 files
- Total processes:
  - mProjectPP : 6,669
  - mDiff & mFitplane : 10,521
  - mBackground : 3,474
  - mAdd & mShrink : 324