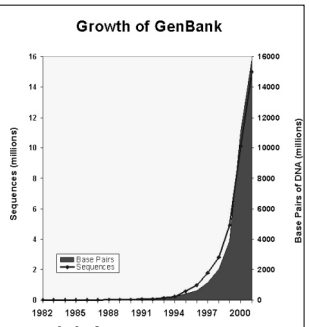
Grid Data Management

Motivation: The Data Problem

- Motivate our discussion with the large physics experiments
  - Laser Interferometer Gravitational Wave Observatory
    - Detect spacetime ripples from blackholes & other sources
    - Generates data at 10 MB per second, just under 1 TB per day
  - Sloan Digital Sky Survey
    - Catalog more stars and galaxies then ever before
    - More than 15 TB of data catalogs
  - Compact Muon Solenoid and ATLAS
    - Detect the Higgs Boson (a fundamental particle)
    - 100 MB per second, about 1 Petabyte per year (per detector)

Technology Drivers

- Internet revolution: 100M+ hosts
  - Collaboration & sharing the norm
- Universal Moore's law: x10^3/10 yrs
  - Sensors as well as computers
- Petascale data tsunami
  - Gating step is analysis
- & our old infrastructure?



114 genomes
735 in progress

Slide compliments of Ian Foster

Really Two Data Problems

- The *amount* of data
  - High-performance tools needed to manage the huge raw volume of data
    - Store it
    - Move it
  - Measure in terabytes, petabytes, and ???
- The *number* of data files
  - High-performance tools needed to manage the huge number of filenames
    - 10^{12} filenames is expected soon
    - Collection of 10^{12} of anything is a lot to handle efficiently

Data Questions on the Grid

Questions for which you want Grid tools to address

- Where are the files I want?
- How to move data/files to where I want?

GridFTP

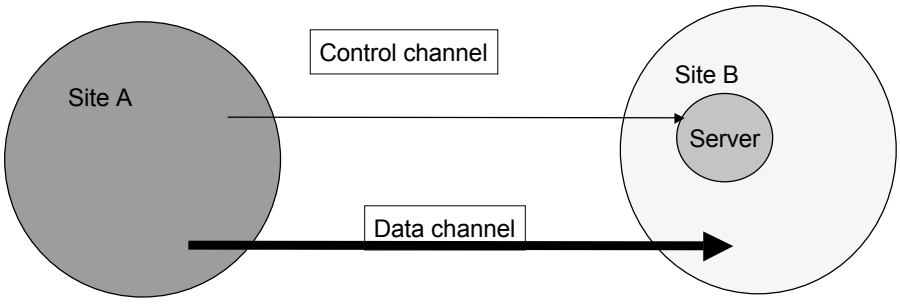
- Extension to well known File Transfer Protocol (FTP)
- http://www.ogf.org/documents/GFD.20.pdf
- Extensions include
  - Strong authentication, encryption via Globus GSI
  - Multiple, parallel data channels
  - Third-party transfers
  - Tunable network & I/O parameters
  - Server side processing, command pipelining

Basic Definitions

- Control Channel
  - TCP link over which commands and responses flow
  - Low bandwidth; encrypted and integrity protected by default
- Data Channel
  - Communication link(s) over which the actual data of interest flows
  - High Bandwidth; authenticated by default; encryption and integrity protection optional

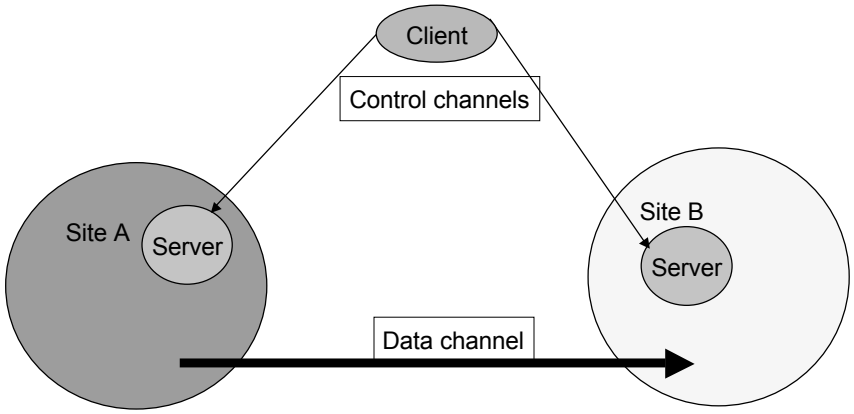
A file transfer with GridFTP

- Control channel can go either way
  - Depends on which end is client, which end is server
- Data channel is still in same direction



Third party transfer

- Controller can be separate from src/dest
- Useful for moving data from storage to compute



globus-url-copy

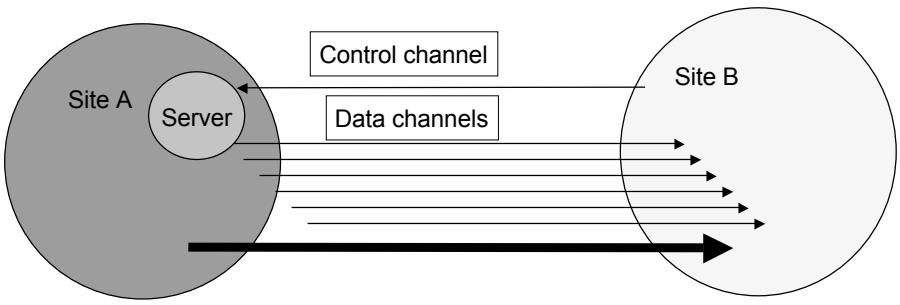
- Globus-url-copy is commandline client for gridftp (and other protocols)
- Already seen this in exercises
- `globus-url-copy gsiftp://gridlab1/home/benc/foo file:///tmp/a`

URLs

- `globus-url-copy gsiftp://gridlab1/home/benc/foo file:///tmp/a`
- Gsiftp – GridFTP server
  - Specify hostname (optionally port)
  - Directory and filename
- File – local filesystem
  - No hostname
  - Just local filename

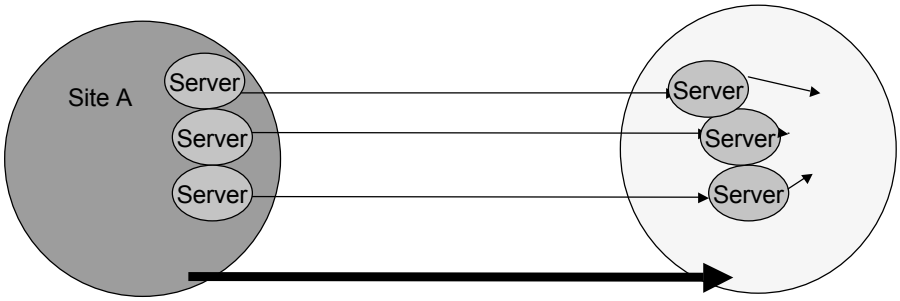
Going fast – parallel streams

- Use several data channels



Going fast – striped transfers

- Use several servers at each end
- Shared storage at each end



Going fast –buffers and windows

- Using large TCP windows

```
$ globus-url-copy -vb -p 4 -tcp-bs 1048576 gsiftp://ldas-
   cit.ligo.caltech.edu:15000/usr1/grid/largefile file:/tmp/largefile
    514392064 bytes      6609.67 KB/sec avg      8639.71 KB/sec inst
```

- Using large memory buffers

```
$ globus-url-copy -vb -p 4 -bs 1048576 -tcp-bs 1048576 gsiftp://ldas-
   cit.ligo.caltech.edu:15000/usr1/grid/largefile file:/tmp/largefile
    523304960 bytes      7300.56 KB/sec avg      9311.99 KB/sec inst
```

- Speed depends on network weather – what else is happening on the network.

Debugging

Use –dbg to see control channel communication

```
$ globus-url-copy -dbg gsiftp://hydra.phys.uwm.edu/tmp/file1 file:/tmp/file1
debug: starting to get gsiftp://hydra.phys.uwm.edu/tmp/file1
debug: connecting to gsiftp://hydra.phys.uwm.edu/tmp/file1
debug: response from gsiftp://hydra.phys.uwm.edu/tmp/file1:
220 hydra.phys.uwm.edu GridFTP Server 1.12 GSSAPI type Globus/GSI wu-2.6.2 (gcc32dbg,
    1069715860-42) ready.

debug: authenticating with gsiftp://hydra.phys.uwm.edu/tmp/file1
debug: response from gsiftp://hydra.phys.uwm.edu/tmp/file1:
230 User skoranda logged in.

debug: sending command:
FEAT

debug: response from gsiftp://hydra.phys.uwm.edu/tmp/file1:
211-Extensions supported:
 REST STREAM
 ESTO
 ERET
 MDTM
 SIZE
 PARALLEL
 DCAU
211 END
<snip>
```


GridFTP clients

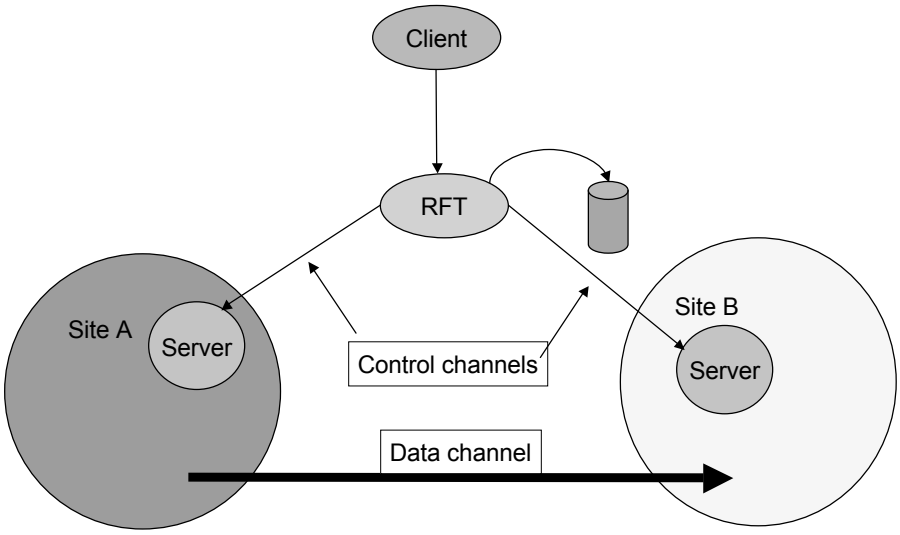
- "Roll your own"
- Add functionality *directly* to your applications
  - Your application find and download its own data?
  - Your application deliver output data files when finished computing?
- Globus Toolkit offers APIs to code against
  - C
  - Java
  - Python

Hints for Experts

To make GridFTP go really fast

- use fast disks/filesystems
  - filesystem should read/write > 30 MB/second
- configure TCP for performance
  - See TCP Tuning Guide at http://www-didc.lbl.gov/TCP-tuning/
- patch your Linux kernel with web100 patch
  - See http://www.web100.org
  - Important work-around for Linux TCP "feature"
- understand your network path

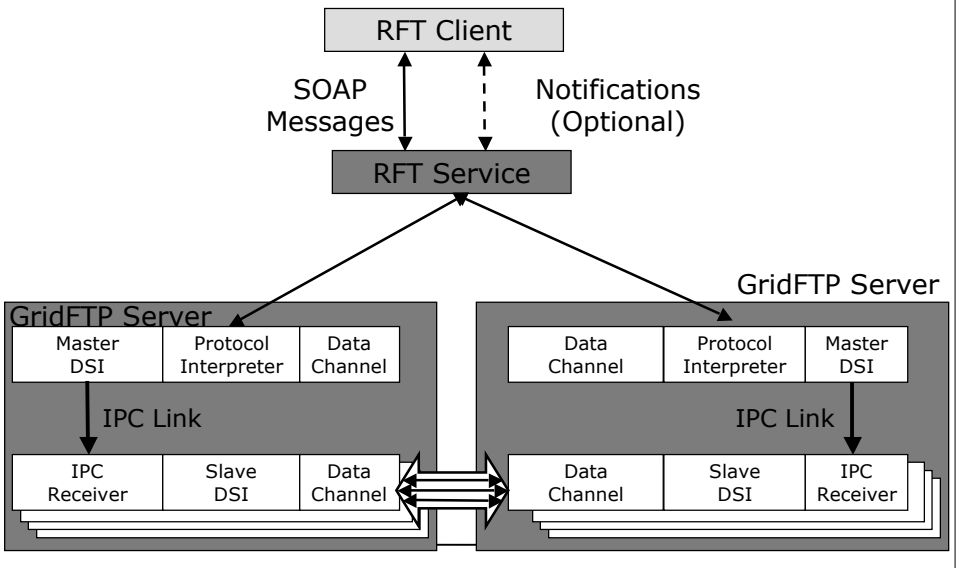
Reliable file transfer



What is RFT ?

- WS-RF compliant High performance data transfer service
  - Soft state.
  - Notifications/Query
- Reliability on top of high performance provided by GridFTP.
  - Fire and Forget.
  - Integrated Automatic Failure Recovery.
    - Network level failures.
    - System level failures etc.

Server Architecture



eXtensible Input/Output (XIO)

- GridFTP server uses xio_read, xio_write, etc..
- XIO uses the concept of a protocol stack
  - By default GridFTP uses TCP with GSI authentication
  - We have demonstrated GridFTP running over UDT (reliable UDP)
- Gives a nice abstraction from the network protocol
- Can be used to add functionality needed on each read and write, such as compression or encryption

Data Questions on the Grid

Questions for which you want Grid tools to address

- Where are the files I want?
- How to move data/files to where I want?

What data/files are where?

- Requirements
  - Catalog 10^8 files and their locations
    - What files are where (possibly at more then one place)
    - Across multiple sites within a Grid
    - Mappings from logical filenames (LFNs) to physical filenames (PFNs) or URLs
  - No single point of failure
    - No central catalog/server to be single point of failure

Globus Replica Location Service

- One solution to this is…
- Globus RLS
  - Maps logical filenames to physical filenames
  - Two components
    - LRC (Local replica catalog)
    - RLI (Replica location index)

Logical and Physical Filenames

- Logical Filenames
  - Names a file with interesting data in it
  - Doesn't refer to location (which host, or where inside a host)
- Physical Filenames
  - Refers to a file on some filesystem somewhere
  - Often use gsiftp:// URLs to specify

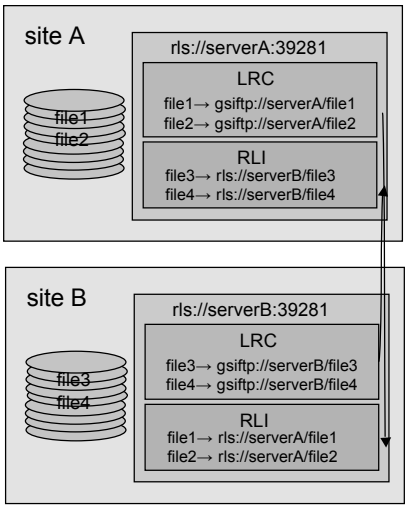
Two catalogs in RLS

- Local Replica Catalog (LRC)
  - Stores mappings from LFNs to PFNs
  - Interaction:
    - Q: Where can I get filename 'experiment_result_1'.
    - A: You can get it from gsiftp://gridlab1/home/benc/r.txt
  - Undesirable to have one of these for whole grid
  - Lots of data
  - Single point of failure

Two catalogs in RLS

- Replica Location Index (RLI)
  - Stores mappings from LFNs to LRCs
  - Interaction:
    - Q: Who can tell me about filename ‘experiment_result_1’.
    - A: You can get more info from the LRC at gridlab1
    - (then go to ask that LRC for more info)
  - Failure of one RLI or LRC doesn’t break everything
  - RLI stores reduced set of information, so can cope with many more mappings

Globus RLS



Globus RLS

- Quick Review
  - LFN → logical filename (think of as simple filename)
  - PFN → physical filename (think of as a URL)
  - LRC → your local catalog of maps from LFNs to PFNs
    - H-R-792845521-16.gwf → gsiftp://dataserver.phys.uwm.edu/LIGO/H-R-792845521-16.gwf
  - RLI → your local catalog of maps from LFNs to LRCs
    - H-R-792845521-16.gwf → LRCs at MIT, PSU, Caltech, and UW-M
  - LRCs inform RLIs about mappings known
  - Find files on your Grid by querying RLI(s) to get LRC(s), then query LRC(s) to get URL(s)

Globus RLS: Server Perspective

- Mappings LFNs → PFNs kept in database
  - Uses generic ODBC interface to talk to any (good) RDBM
  - MySQL, PostgreSQL, Oracle, DB2,...
  - All RDBM details hidden from administrator and user
    - well, not quite
    - RDBM may need to be "tuned" for performance
    - but one can start off knowing very little about RDBMs

Globus RLS: Server Perspective

Mappings LFNs → LRCs stored in 1 of 2 ways

- table in database
  - full, complete listing from LRCs that update your RLI
  - requires each LRC to send your RLI full, complete list
    - as number of LFNs in catalog grows, this becomes substantial
    - 10^8 filenames at 64 bytes per filename ~ 6 GB
- in memory in a special hash called Bloom filter
  - 10^8 filenames stored in as little as 256 MB
    - easy for LRC to create Bloom filter and send over network to RLIs
  - can cause RLI to lie when asked if knows about a LFN
    - only false-positives
    - tunable error rate
    - acceptable in many contexts
  - Wild carding not possible with Bloom Filters

RLS command line tools

- globus-rls-admin
  - administrative tasks
    - ping server
    - connect RLIs and LRCs together
- globus-rls-cli
  - end user tasks
    - query LRC and RLI
    - add mappings to LRC

Globus RLS: Client Perspective

Two ways for clients to interact with RLS Server

- globus-rls-cli simple command-line tool
  - query
  - create new mappings
- "roll your own" client by coding against API
  - Java
  - C
  - Python

Globus-rls-cli

Simple query to LRC to find a PFN for LFN

- Note more then 1 PFN may be returned

```
$ globus-rls-cli query lrc lfn H-R-714024224-16.gwf rls://dataserver:39281
  H-R-714024224-16.gwf: file://localhost/netdata/s001/S1/R/H/714023808-
   714029599/H-R-714024224-16.gwf
  H-R-714024224-16.gwf: file://medusa-
   slave001.medusa.phys.uwm.edu/data/S1/R/H/714023808-714029599/H-R-
   714024224-16.gwf
  H-R-714024224-16.gwf:
   gsiftp://dataserver.phys.uwm.edu:15000/data/gsiftp_root/cluster_storage/
   data/s001/S1/R/H/714023808-714029599/H-R-714024224-16.gwf
```

- Server and client sane if LFN not found

```
$ globus-rls-cli query lrc lfn "foo" rls://dataserver
LFN doesn't exist: foo
$ echo $?
1
```

Globus-rls-cli

Wildcard searches of LRC supported

- probably a good idea to quote LFN wildcard expression

```
$ globus-rls-cli query wildcard lrc lfn "H-R-7140242*-16.gwf"
   rls://dataserver:39281
  H-R-714024208-16.gwf:
   gsiftp://dataserver.phys.uwm.edu:15000/data/gsiftp_root/cluster_stor
   age/data/s001/S1/R/H/714023808-714029599/H-R-714024208-16.gwf
  H-R-714024224-16.gwf:
   gsiftp://dataserver.phys.uwm.edu:15000/data/gsiftp_root/cluster_stor
   age/data/s001/S1/R/H/714023808-714029599/H-R-714024224-16.gwf
```


Globus-rls-cli

Bulk queries also supported

- obtain PFNs for more then one LFN at a time

```
$ globus-rls-cli bulk query lrc lfn H-R-714024224-16.gwf
   H-R-714024320-16.gwf rls://dataserver
  H-R-714024320-16.gwf:
   gsiftp://dataserver.phys.uwm.edu:15000/data/gsiftp_root/
   cluster_storage/data/s001/S1/R/H/714023808-714029599/H-
   R-714024320-16.gwf
  H-R-714024224-16.gwf:
   gsiftp://dataserver.phys.uwm.edu:15000/data/gsiftp_root/
   cluster_storage/data/s001/S1/R/H/714023808-714029599/H-
   R-714024224-16.gwf
```

Globus-rls-cli

Simple query to RLI to locate a LFN to LRC map

- then query that LRC for the PFN

```
$ globus-rls-cli query rli lfn H-R-714024224-16.gwf
   rls://dataserver
  H-R-714024224-16.gwf: rls://ldas-cit.ligo.caltech.edu:39281

$ globus-rls-cli query lrc lfn H-R-714024224-16.gwf
   rls://ldas-cit.ligo.caltech.edu:39281
  H-R-714024224-16.gwf: gsiftp://ldas-
   cit.ligo.caltech.edu:15000/archive/S1/L0/LHO/H-R-7140/H-R-
   714024224-16.gwf
```


Globus-rls-cli

- Bulk queries to RLI also supported

```
$ globus-rls-cli bulk query rli lfn H-R-714024224-16.gwf H-R-
   714024320-16.gwf rls://dataserver
  H-R-714024320-16.gwf: rls://ldas-cit.ligo.caltech.edu:39281
  H-R-714024224-16.gwf: rls://ldas-cit.ligo.caltech.edu:39281
```

- Wildcard queries to RLI may not be supported!
  - no wildcards when using Bloom filter updates

```
$ globus-rls-cli query wildcard rli lfn "H-R-7140242*-
   16.gwf" rls://dataserver
Operation is unsupported: Wildcard searches with Bloom
   filters
```

Globus-rls-cli

Create new LFN → PFN mappings

- use `create` to create 1st mapping for a LFN

```
$ globus-rls-cli create file1 gsiftp://dataserver/file1
  rls://dataserver
```

- use `add` to add more mappings for a LFN

```
$ globus-rls-cli add file1 file://dataserver/file1
  rls://dataserver
```

- use `delete` to remove a mapping for a LFN
  - when last mapping is deleted for a LFN the LFN is also deleted
  - cannot have LFN in LRC without a mapping

```
$ globus-rls-cli delete file1 file://file1 rls://dataserver
```


Globus-rls-cli

LRC can also store attributes about LFN and PFNs

- size of LFN in bytes?
- md5 checksum for a LFN?
- ranking for a PFN or URL?
- extensible...you choose attributes to create and add
- can search catalog on the attributes
- attributes limited to
  - strings
  - integers
  - floating point (double)
  - date/time

Globus-rls-cli

- Create attribute first then add values for LFNs

```
$ globus-rls-cli attribute define md5checksum lfn string
   rls://dataserver
$ globus-rls-cli attribute add file1 md5checksum lfn
   string 42947c86b8a08f067b178d56a77b2650 rls://dataserver
```

- Then query on the attribute

```
$ globus-rls-cli attribute query file1 md5checksum lfn
   rls://dataserver
  md5checksum: string: 42947c86b8a08f067b178d56a77b2650
```


Bloom filters

- LRC-to-RLI flow can happen in two ways:
  - LRC sends list of all its LFNs (but not PFNs) to the RLI. RLI stores whole list.
    - Answer accurately: “Yes I know” / “No I don’t know”
    - Expensive to move and store large list
  - Bloom filters
    - LRC generates a Bloom filter of all of its LFNs
    - Bloom filter is a bitmap that is much smaller than whole list of LFNs
    - Answers less accurately: “Maybe I know” / “No I don’t know”. Might end up querying LRCs unnecessarily (but we won’t ever get wrong answers)
    - can’t do a wildcard search

Related Work

- Storage Resource Manager (SRM)
  - Equivalent of a job scheduler for storage; allocates space, makes sure it doesn't get swapped out before you are done (pinning); handles staging to and from tape
  - http://sdm.lbl.gov/indexproj.php?ProjectID=SRM
- dCache
  - provide a system for storing and retrieving huge amounts of data, distributed among a large number of heterogenous server nodes, under a single virtual filesystem tree with a variety of standard access methods.
  - http://www.dcache.org/

Related Work

- Globus Metadata Catalog
  - a stand-alone metadata catalog service with an OGSA service interface. The metadata catalog associates application-specific descriptions with data files, tables, or objects. These descriptions, which are encoded in structured ways defined by "schema" or community standards, make it easier for users and applications to locate data relevant to specific problems.
    - "I want the temperature, barometric pressure, and CO2 concentrations for this geographic area"
  - http://www.globus.org/grid_software/data/mcs.php

Related Work

- Stork
  - Cross between RFT and Condor DAGMAN
  - make data placement activities "first class citizens" in the Grid just like the computational jobs. They will be queued, scheduled, monitored, managed, and even check-pointed. More importantly, it will be made sure that they complete successfully and without any human interaction.
  - http://www.cs.wisc.edu/condor/stork/
- Storage Resource Broker
  - supports shared collections that can be distributed across multiple organizations and heterogeneous storage systems. The SRB can be used as a Data Grid Management System (DGMS) that provides a hierarchical logical namespace to manage the organization of data (usually files).
  - http://www.sdsc.edu/srb/index.php/Main_Page

Credits

Bill Allcock allcock@mcs.anl.gov

based on slides from
Ben Clifford benc@ci.uchicago.edu
Scott Koranda skoranda@uwm.edu