

Grid-based Digital Libraries and Cheshire3

Ray R. Larson

University of California,
Berkeley

School of Information

In collaboration with

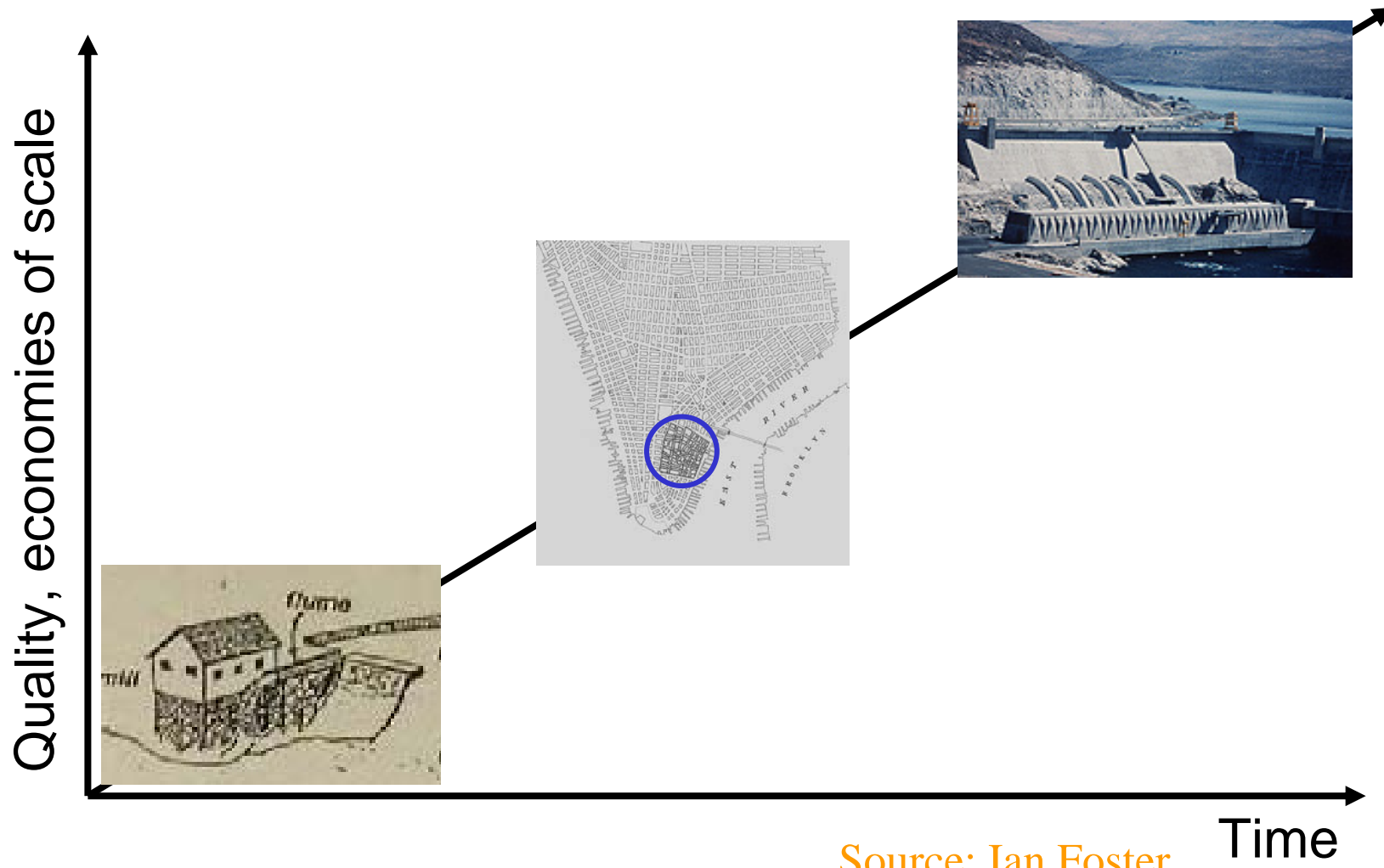
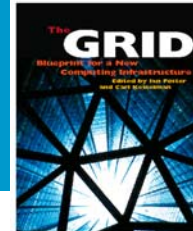
Robert Sanderson

University of Liverpool

Department of Computer Science

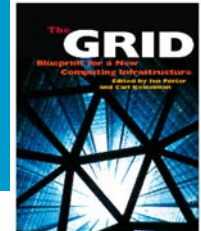
- The Grid and Digital Libraries
 - Grid Architecture
 - Grid IR Issues
 - Grid Environment for DL
- Cheshire3:
 - Overview
 - Cheshire3 Architecture
 - Distributed Workflows
 - Grid Experiments

The Grid: On-Demand Access to Electricity



Source: Ian Foster

By Analogy, A Computing Grid



- Decouples production and consumption
 - Enable on-demand access
 - Achieve economies of scale
 - Enhance consumer flexibility
 - Enable new devices
- On a variety of scales
 - Department
 - Campus
 - Enterprise
 - Internet

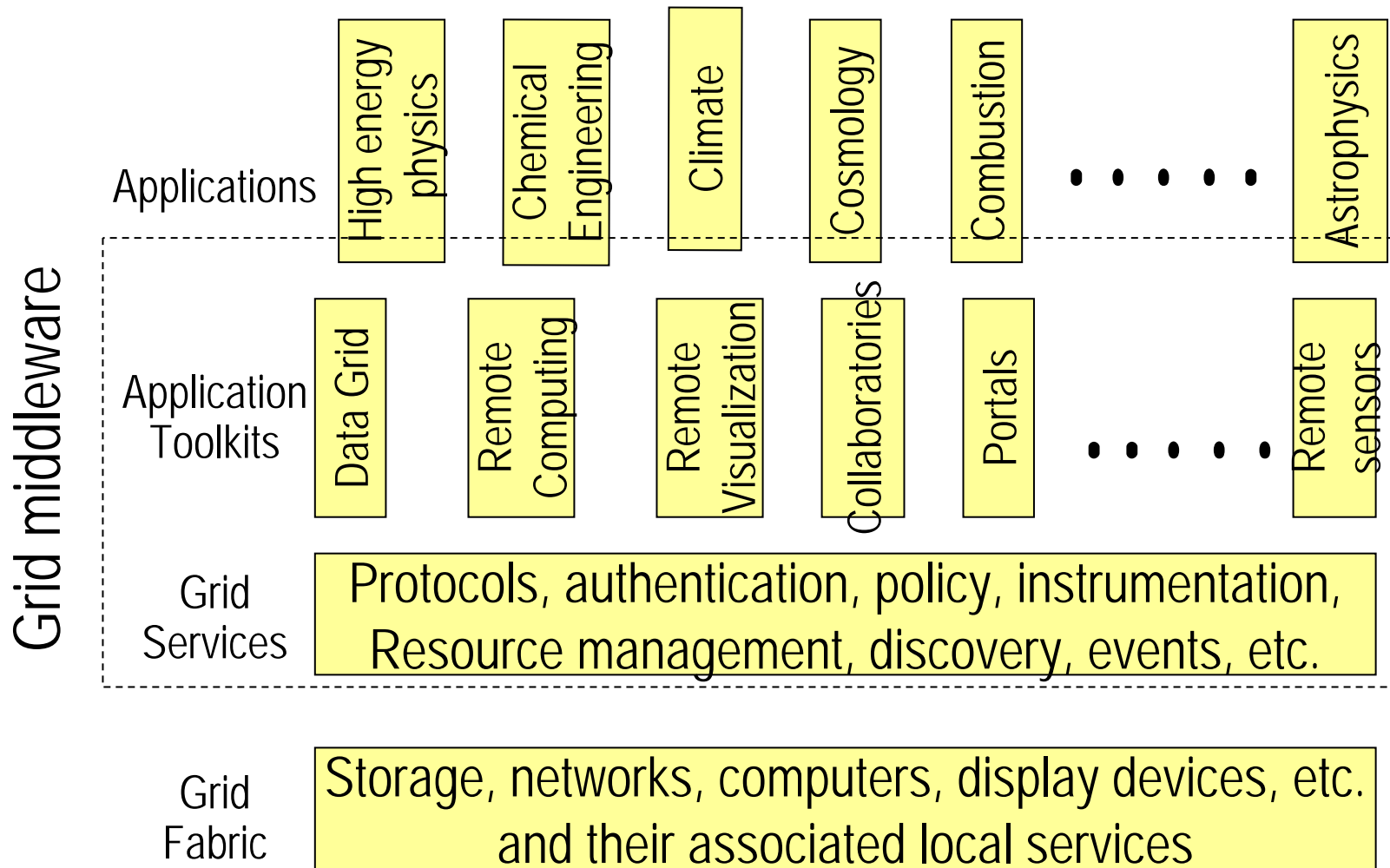
Source: Ian Foster

What is the Grid?



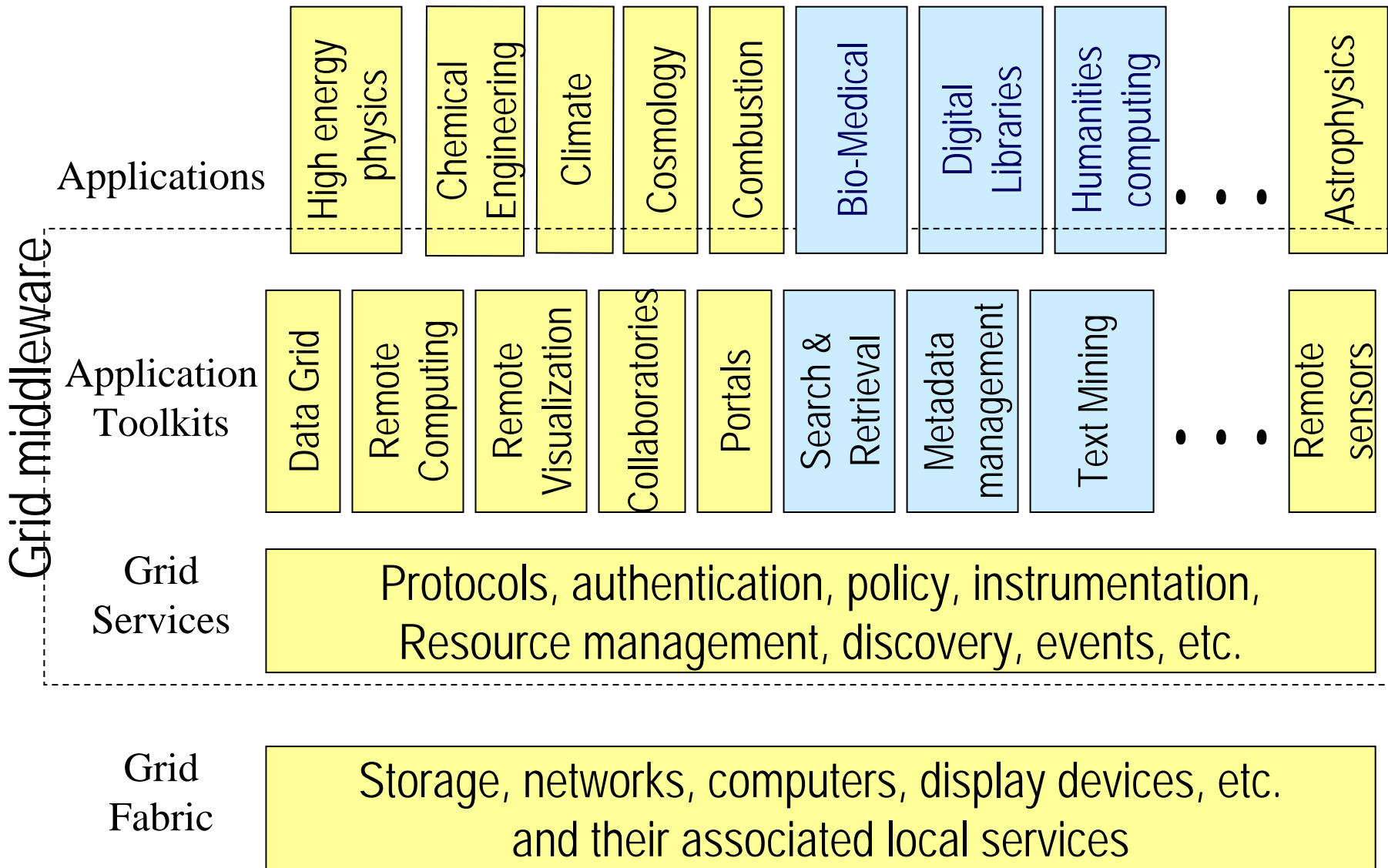
“The short answer is that, whereas the Web is a service for sharing information over the Internet, the Grid is a service for sharing computer power and data storage capacity over the Internet. The Grid goes well beyond simple communication between computers, and aims ultimately to turn the global network of computers into one vast computational resource.”

Source: The Global Grid Forum



- Applications and data that are NOT for scientific research?
- Things like:

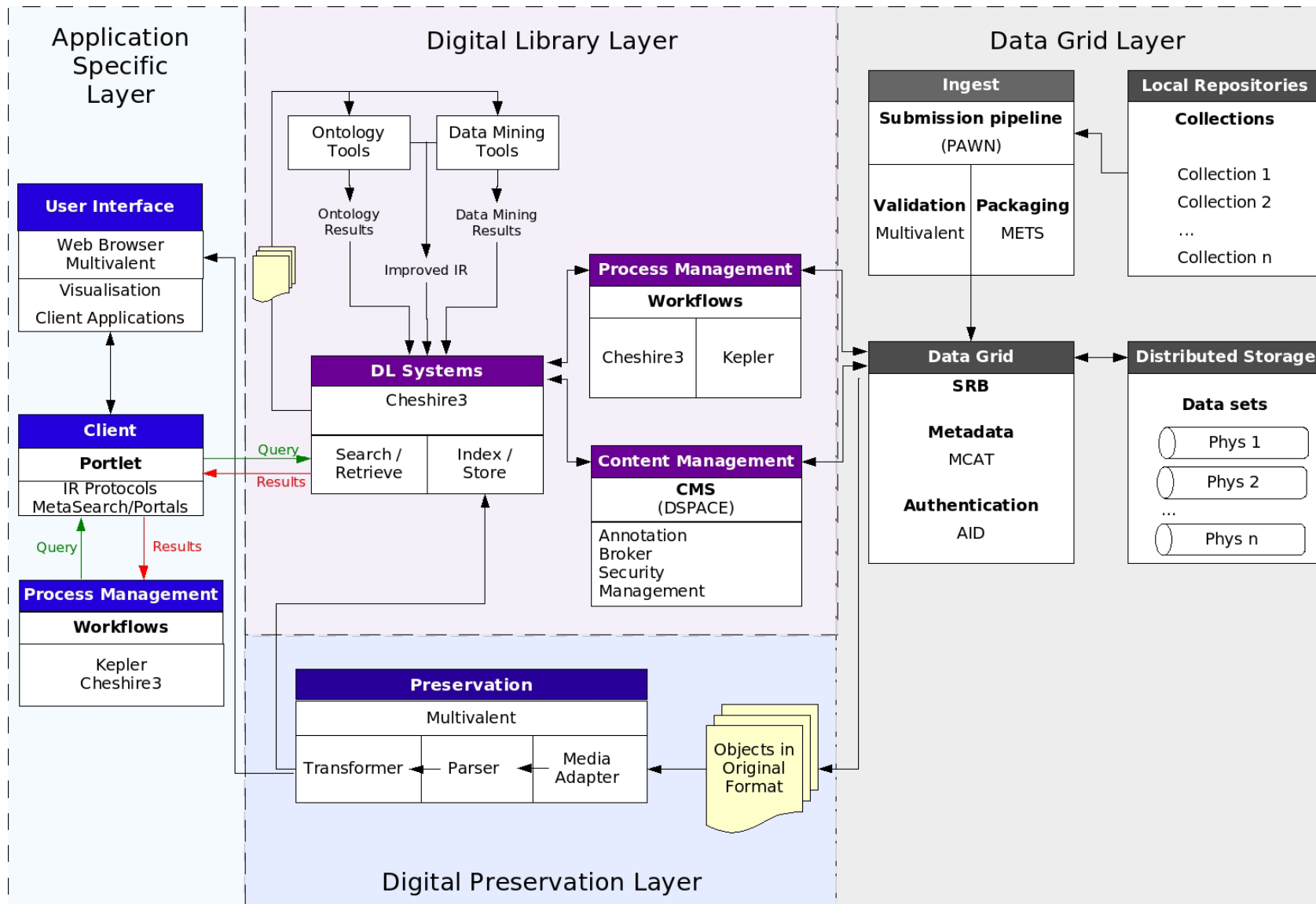
Digital Libraries?

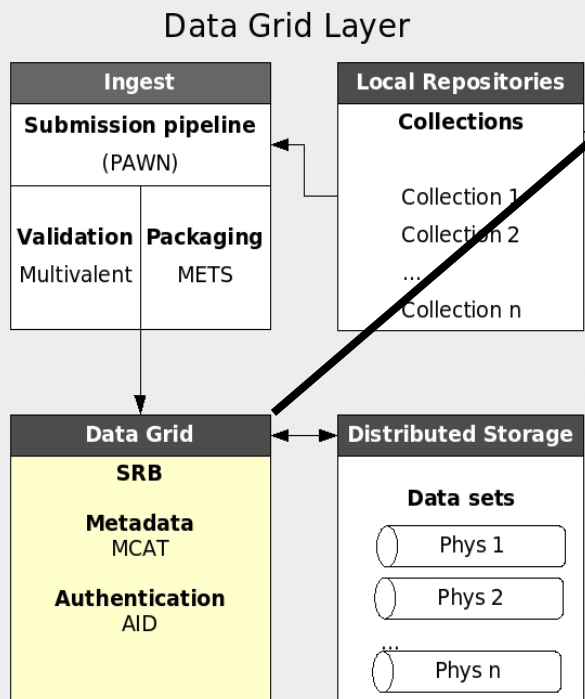


- Large-scale distributed storage requirements and technologies
- Organizing distributed digital collections
- Shared Metadata – standards and requirements
- Managing distributed digital collections
- Security and access control
- Collection Replication and backup
- **Distributed Information Retrieval support and algorithms**

- Want to preserve the same retrieval performance (precision/recall) while hopefully increasing efficiency (i.e. speed)
- Very large-scale distribution of resources is (still) a challenge for sub-second retrieval
- Different from most other typical Grid processes, IR is potentially less computing intensive and more data intensive
- In many ways Grid IR replicates the process (and problems) of metasearch or distributed search
- We have developed the Cheshire3 system to evaluate and manage these issues. The Cheshire3 system is actually one component in a larger Grid-based environment

Cheshire3 Environment





SRB: Storage Resource Broker

DataGrid system for storing Large amounts of data.

Developed at San Diego Supercomputer Center

Advantages:

- Replication
- Storage Resource Abstraction
- Logical identifiers vs 'physical' identifiers
- Mountable as a filesystem

The SRB is emerging as the de-facto standard for data-grid applications, and is already in use by:

- The World University Network
- The Biomedical Informations Research Network (BIRN)
- The UK eScience Centre (CCLRC)
- The National Partnership for Advanced Computational Infrastructure (NPACI)
- NASA information power grid

<http://www.sdsc.edu/srb/>

- Digital Library Content management systems such as Dspace and Fedora, are currently being extended to make use of the SRB for data grid storage
- This will ensure their collections can in future be of virtually unlimited size, and be stored, replicated, and accessible via federated grid technologies
- By supporting the SRB, we have ensured that the Cheshire framework will be able to integrate with these systems, thereby facilitating digital content ingestion, resource discovery, content management, dissemination, and preservation, within a data-grid environment

Kepler/Ptolemy

Process Management

Workflows

Cheshire3

Kepler

Workflow processing environment developed at UC Berkeley (Ptolemy) and SDSC (Kepler) plus others including LLNL, UCSD and University of Zurich.

Director/Actor model:

Actors perform tasks together as directed.

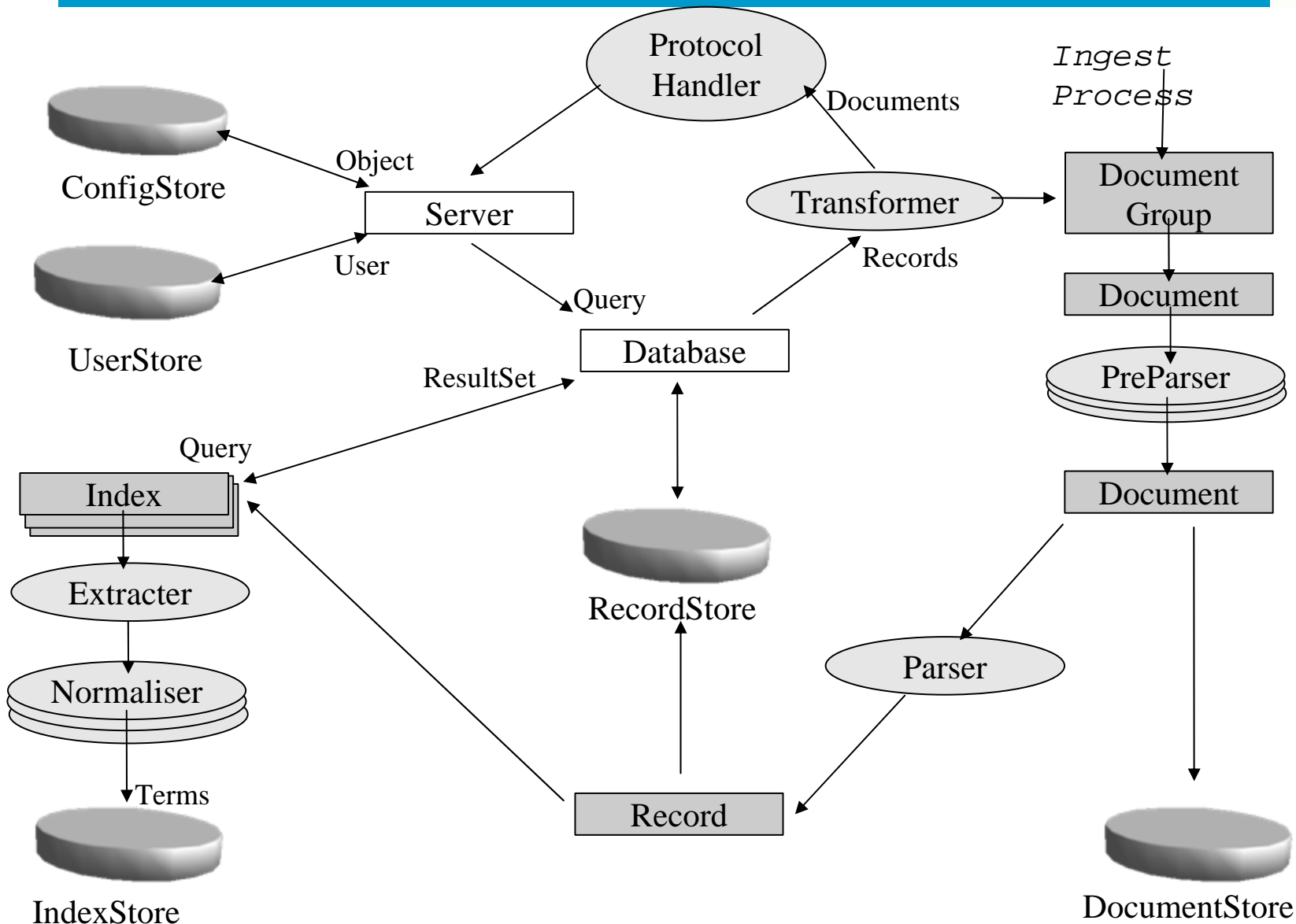
- Workflow environments, such as Kepler, are designed to allow researchers to design and execute flexible processing sequences for complex data analysis
- They provide a Graphical User Interface to allow any level of user from a variety of disciplines to design these workflows in a drag-and-drop manner
- This provides a platform can integrate text mining techniques and methodologies, either as part of an internal Cheshire workflow, or as external workflow configured using a Kepler
<http://kepler-project.org/>

- The Cheshire system is being used in the UK National Text Mining Centre (NaCTeM) as a primary means of integrating information retrieval systems with text mining and data analysis systems
- NARA Prototype which is demonstrating use of the Cheshire3 environment for indexing and retrieval in a preservation environment. Currently we have a web crawl of all information related to the Columbia Shuttle disaster
- NSDL Analysis to analyse 200GB of web-crawled data from the NSDL (National Science Digital Library) and analyse each document for grade level based on vocabulary. We are using LSI and Cluster analysis to categorize the crawled documents
- CURL Data -- 45 Million records of library bibliographic data from major research libraries in the UK

- Cheshire was originally created at UC Berkeley and more recently co-developed at the University of Liverpool. The system itself is widely used in the United Kingdom for production digital library services including:
 - Archives Hub
 - JISC Information Environment Service Registry
 - Resource Discovery Network
 - British Library ISTC service
- The Cheshire system has recently gone through a complete redesign into its current incarnation, Cheshire3 enabling Grid-based IR over the Data Grid

- XML Information Retrieval Engine
 - 3rd Generation of the UC Berkeley Cheshire system, as co-developed at the University of Liverpool
 - Uses Python for flexibility and extensibility, but uses C/C++ based libraries for processing speed
 - Standards based: XML, XSLT, CQL, SRW/U, Z39.50, OAI to name a few
 - Grid capable. Uses distributed configuration files, workflow definitions and PVM or MPI to scale from one machine to thousands of parallel nodes
 - Free and Open Source Software
 - <http://www.cheshire3.org/>

Cheshire3 Object Model



Each non Data Object has an XML configuration.

- Common base schema with extensions as needed.

Configurations can be treated as a Record.

- Store them in regular RecordStores
- Access/Distribute them via regular IR protocols
- (Requires a 'bootstrap' to find the configuration for the configStore)

Each object has a 'pseudo-unique' identifier.

- Unique within the current context (server, database, etc)
- Can re-apply identifiers at a lower level

Workflows are objects in all of the above ways

Cheshire3 workflows are a simple and nonstandard XML definition

Intentional:

- The workflows are specific to the Cheshire3 architecture
- Also dependent on the architecture
- They replace lines of boring code required for every new database
- Most importantly, they replace lines of code in distributed processing

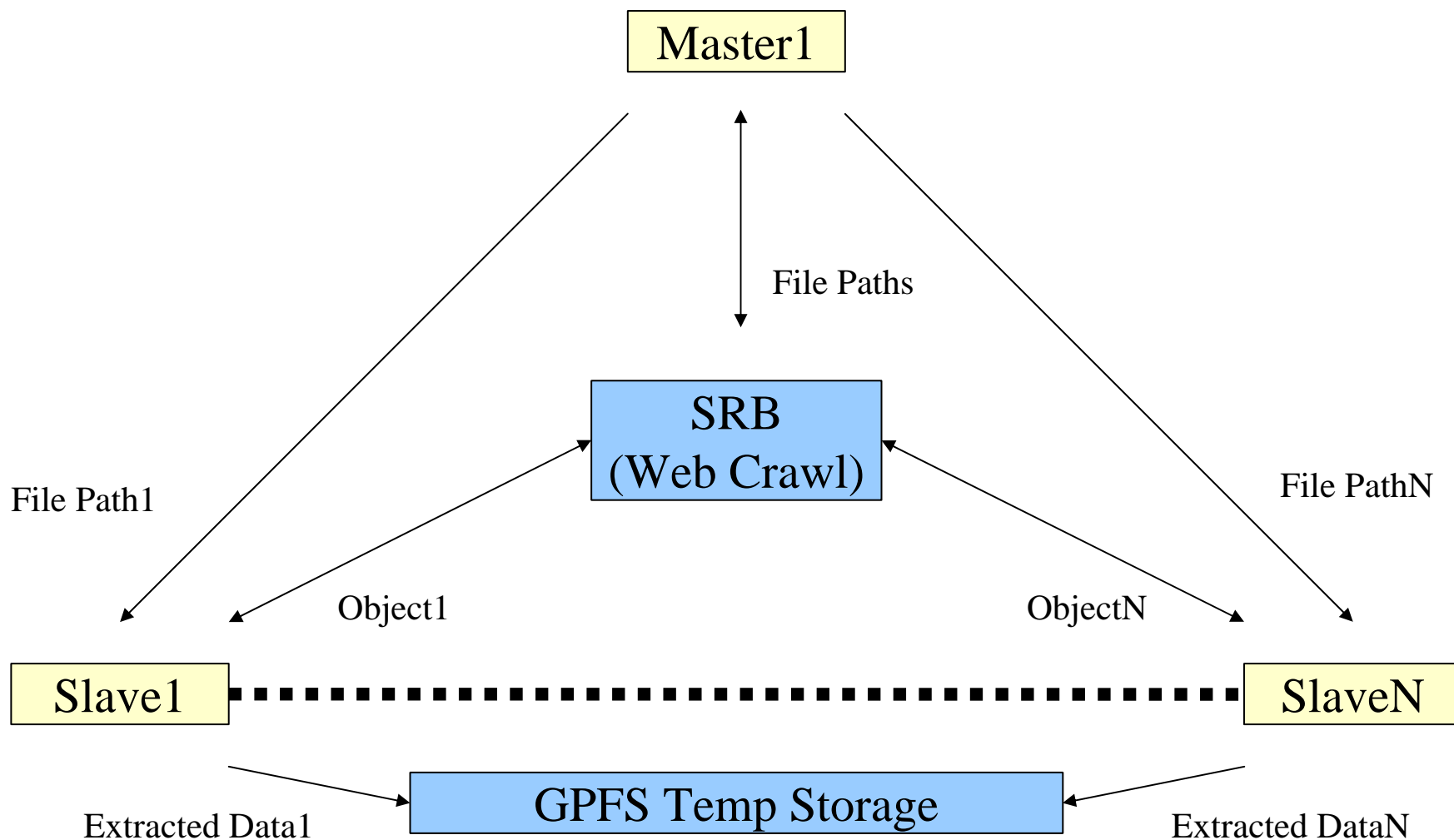
- Need to be easy to understand
- Need to be easy to create

How do workflows help us in massively parallel processing?

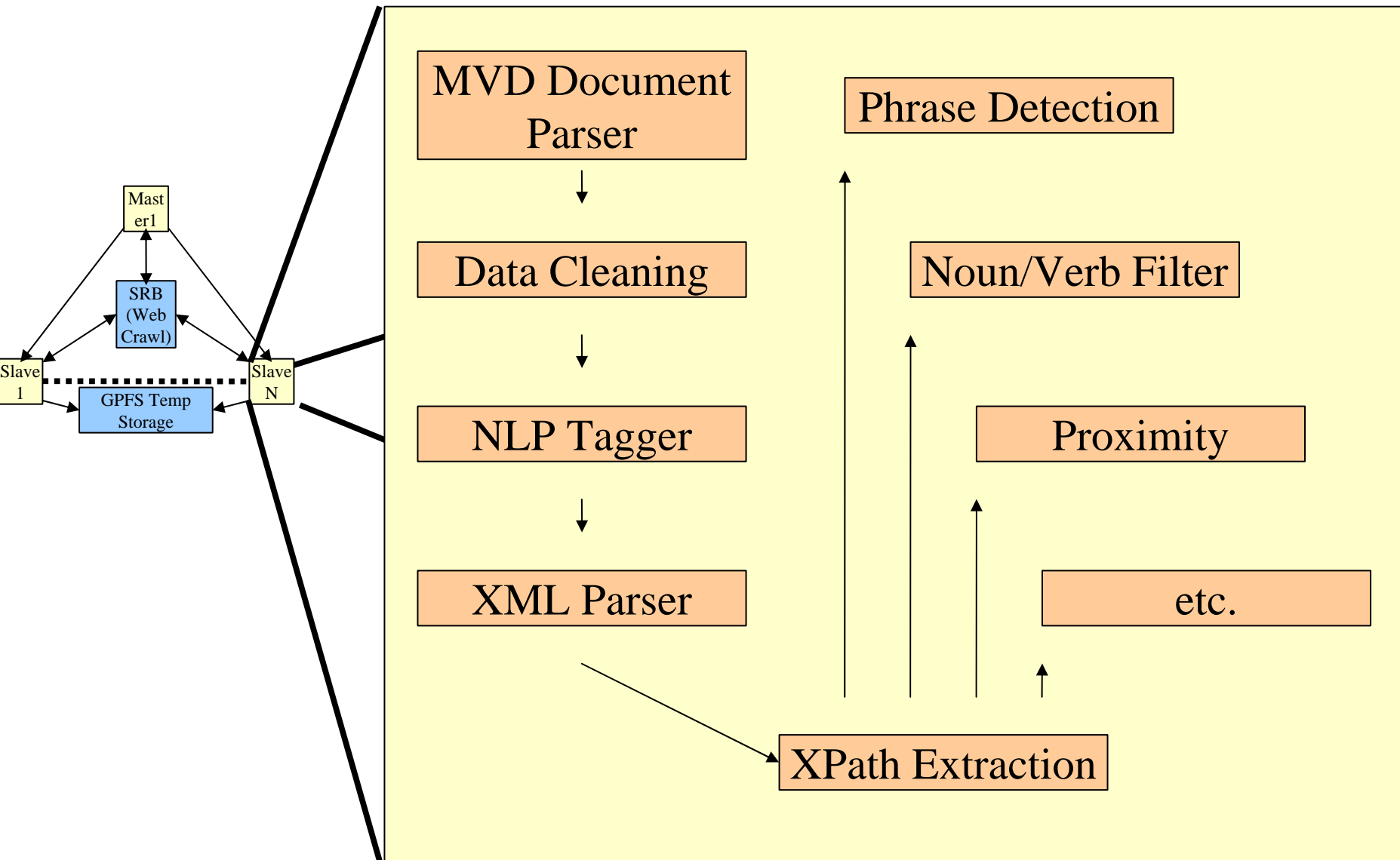
- Each node in the cluster instantiates the configured architecture, potentially through a single ConfigStore
- Master nodes then run a high level workflow to distribute the processing amongst Slave nodes by reference to a subsidiary workflow
- As object interaction is well defined in the model, the result of a workflow is equally well defined. This allows for the easy chaining of workflows, either locally or spread throughout the cluster

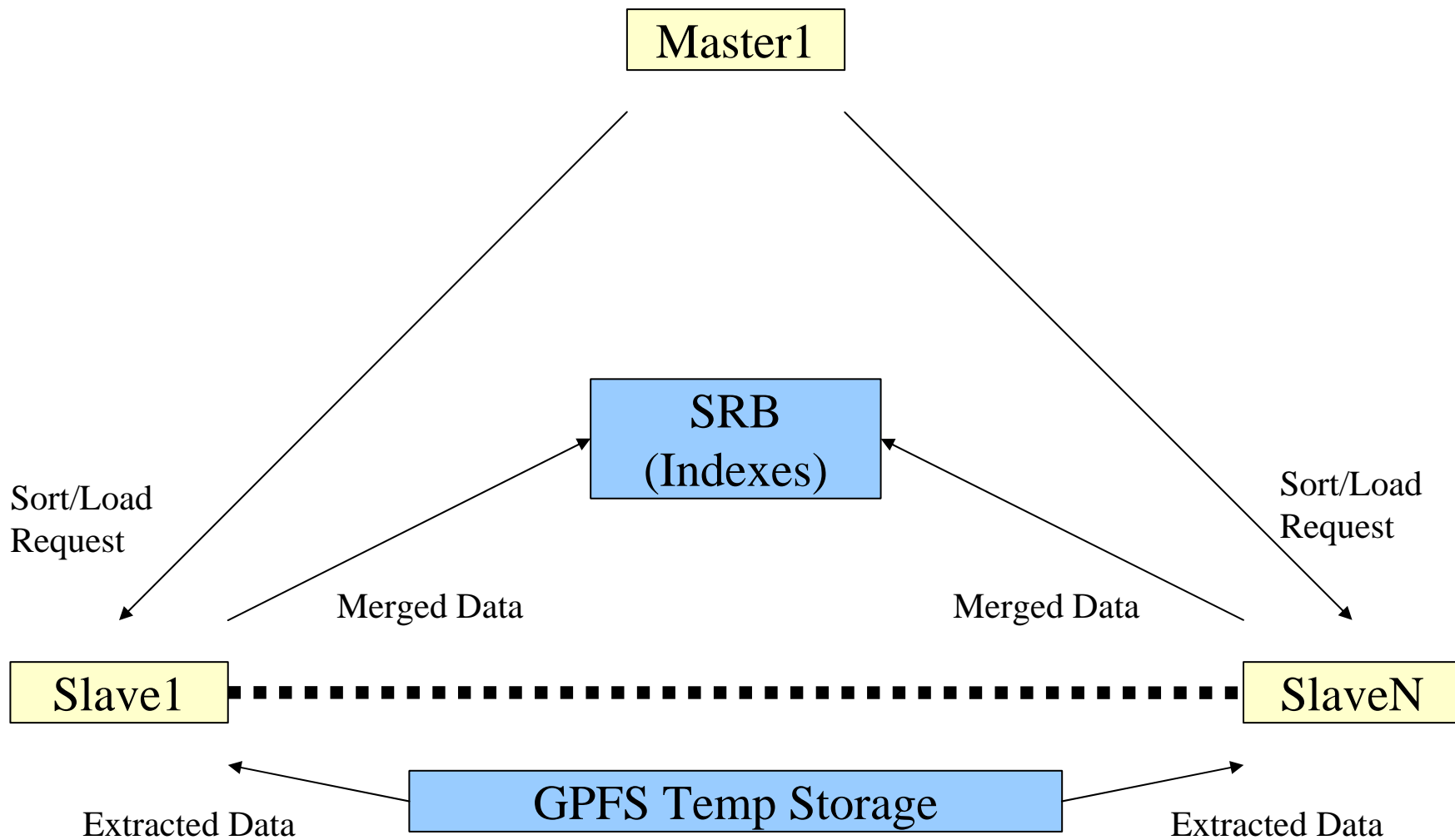
- We are continuing work with SDSC to run evaluations using the TeraGrid through two “small” grants for 30000 CPU hours each
 - SDSC's TeraGrid cluster currently consists of 256 IBM cluster nodes, each with dual 1.5 GHz Intel® Itanium® 2 processors, for a peak performance of 3.1 teraflops. The nodes are equipped with four gigabytes (GBs) of physical memory per node. The cluster is running SuSE Linux and is using Myricom's Myrinet cluster interconnect network
- Large-scale test collections now include MEDLINE, NSDL, the NARA preservation prototype, and the CURL bibliographic data, and we hope to use CiteSeer and the “million books” collections of the Internet Archive
- Using 100 machines, we processed 1.8 million Medline records at a sustained rate of 15,385 per second. With all 256 machines, taking into account additional process management overhead, we could index the entire 16 million record collection in around 7 minutes.
- Using 32 machines, we processed 16 million bibliographic records at a rate of 35,700 records per second. This equates to real time searching of the Library of Congress.

- Two bottlenecks in processing became apparent, even using data distribution, fetching from the SRB and writing to a single recordStore.
- While the cluster has fibre internally, we could only manage 1Mb/second download from the SRB. For simple indexing without NLP, this is a limiting factor.
- In order to maintain sequential numeric record identifiers (eg for compression), a single task was dedicated to writing data to disk. As the 'disk' was a parallel network file system, this also proved to be an I/O bottleneck.

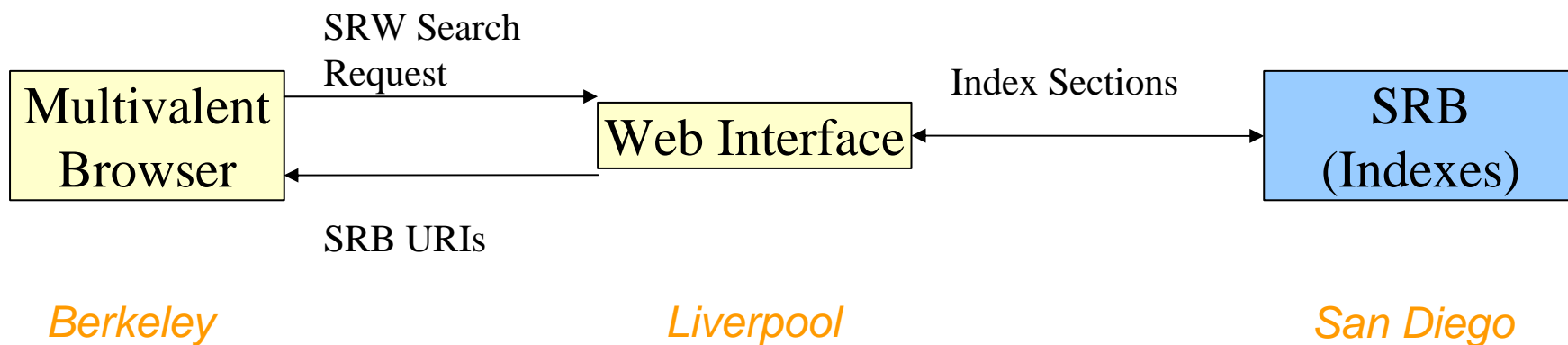


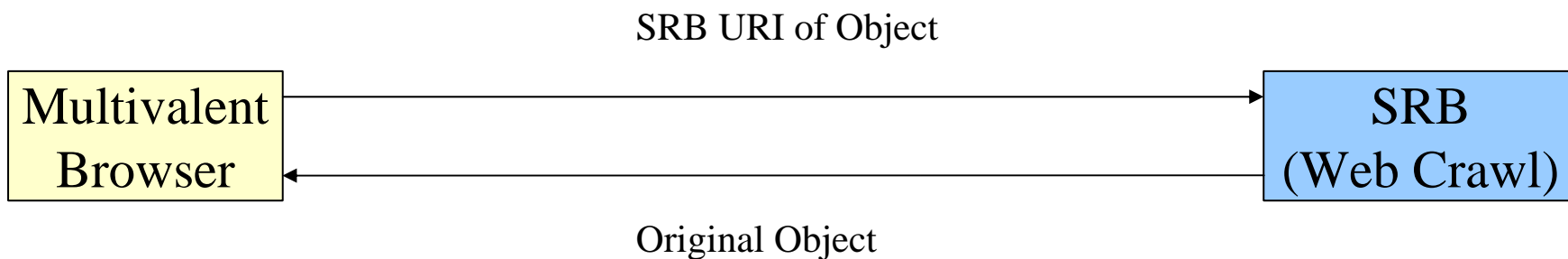
Teragrid Indexing: Slave





In order to locate matching records, the web interface retrieves the relevant chunks of index from the SRB on demand.





- Indexing and IR work very well in the Grid environment, with the expected scaling behavior for multiple processes
- We are working on other large-scale indexing projects (such as the NSDL) and will also be running evaluations of retrieval performance using IR test collections such as the TREC “Terabyte track” collection



Available via <http://www.cheshire3.org>