A Data Storage Language for the Rebels and Misfits

Arun Jagatheesan
San Diego Supercomputer Center (SDSC)
University of California, San Diego

Grid Physics Network
(GriPhyN)

University of Florida

San Diego Supercomputer Center, University of California at San Diego

Talk Outline

- Problem Situation
- Problem
- Solution
  - Concepts
  - If or Why different?
- Status

Boring – Isn't this supposed to be a fun session?
Hint: count the times he will use the word "Actually"

Problem Situation

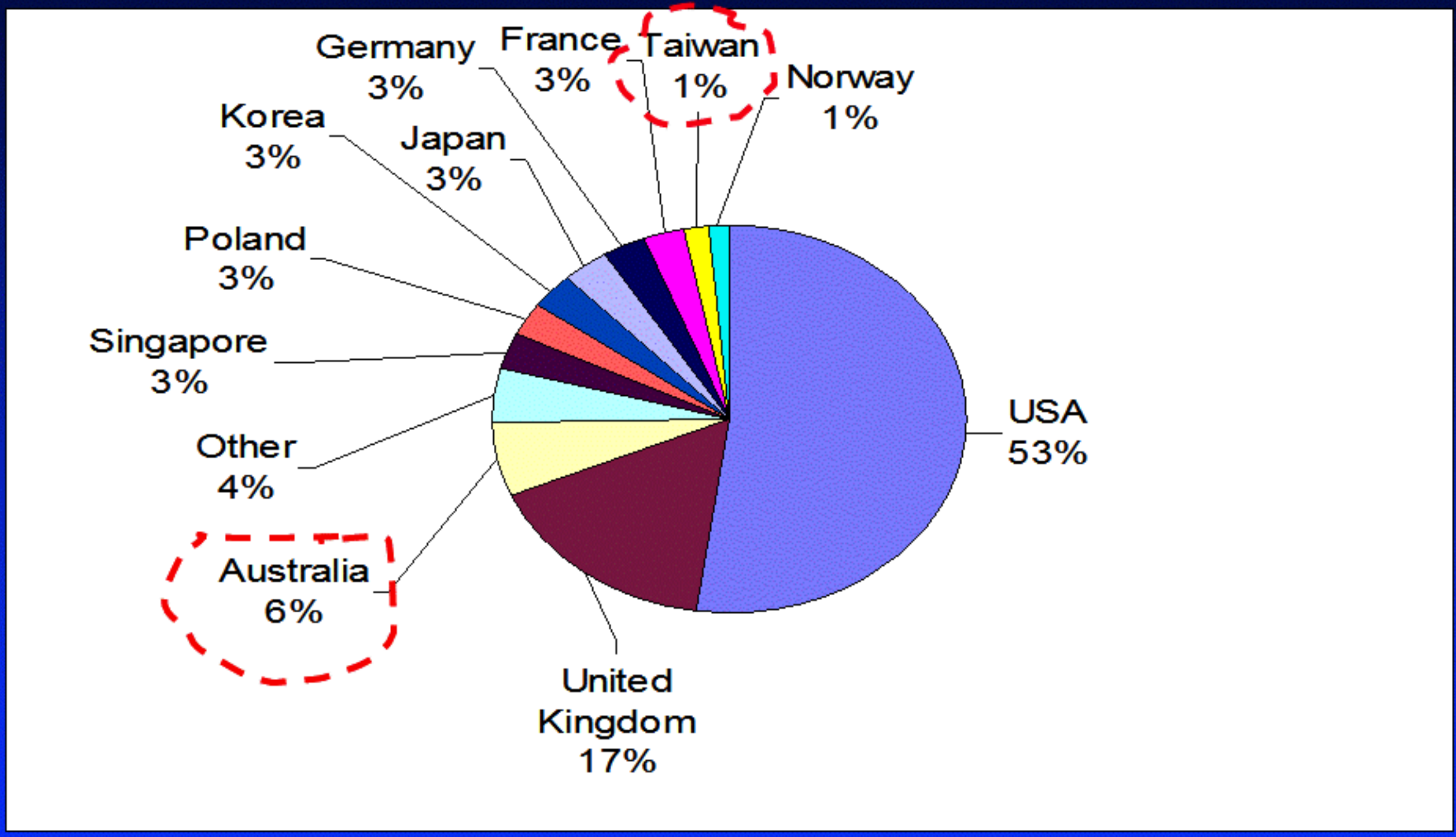
SDSC:

- Petabytes of storage
- Millions of files
- Distributed in a data grid

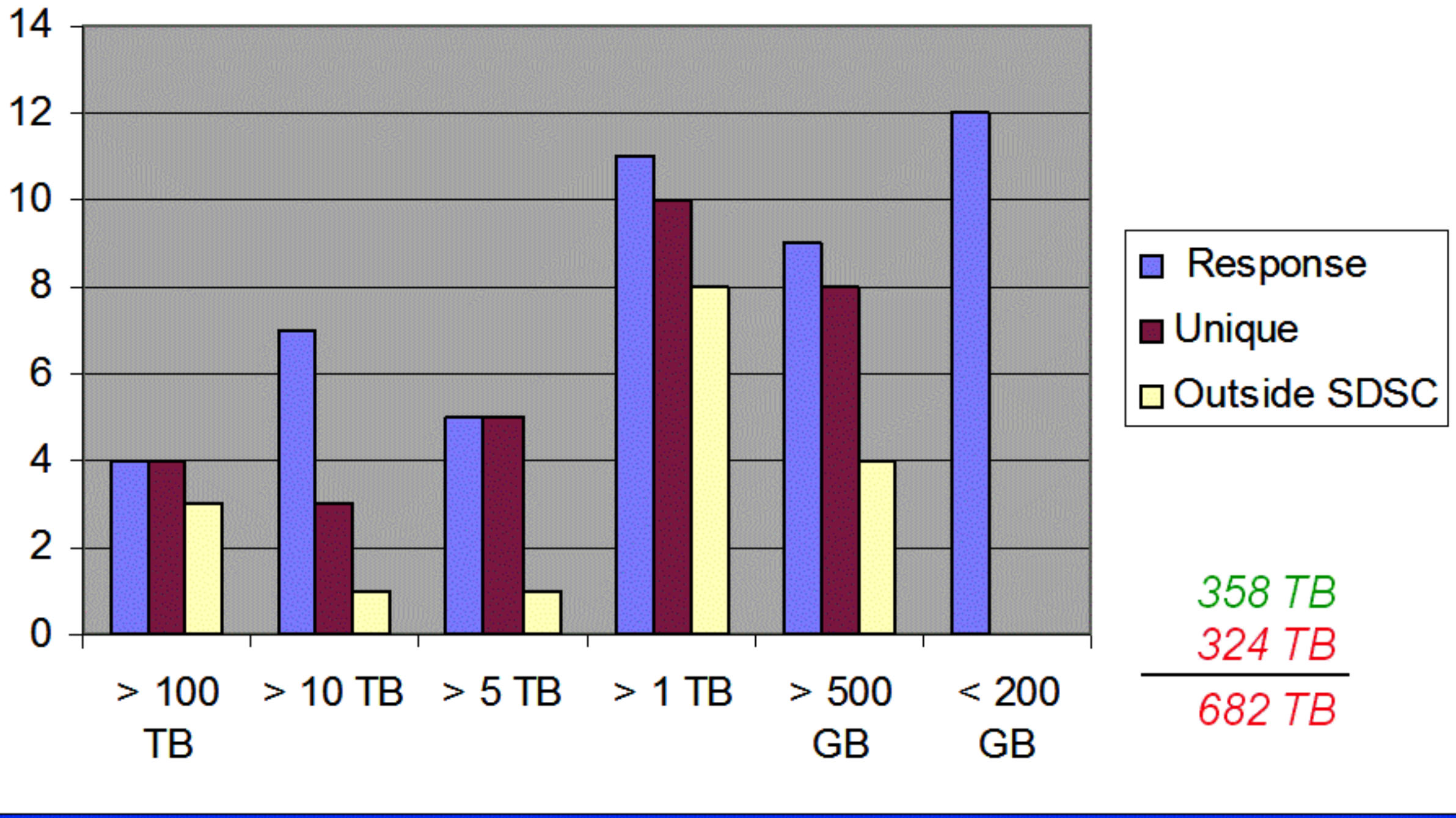
We believe our Data Storage Software already has:

- Infrastructure Independence (more than location independence)
- Data virtualization (more than storage virtualization)

Countries actively using SDSC SRB



Total data brokered by SDSC SRB



SDSC SRB User Community (Major US)

- BaBar, Stanford Linear Accelerator Center (SLAC)
- California Digital Library (CDL)
- Center for Integrated Space Weather Modeling (CISM)
- CVC, Visualization Portal
- LDC Data Storage
- NIH Bio Informatics Research Network (BIRN)
- NSF Southern California Earthquake Center (SCEC)
- National Archives and Records Administration (NARA)
- National Aeronautics and Space Administration Centers (NASA)
- National Virtual Observatory (NVO)
- Npackage, NSF Middleware Initiative (NMI)
- National Science Digital Library (NSDL)
- National Optical Astronomy Observatory (NOAO)
- ROADNet
- Purdue University
- SCCOOS, USA
- Scientific Rich Media Archive
- Salk Institute
- Strand Map Service, USA
- UC Berkeley Library
- UCSD Library
- University of Houston
- Persistent Archives Test bed
- University of Wisconsin, Madison
- WebBase, Stanford University
- Yale University Library

SDSC SRB User Community

- Academia Sinica, Taiwan
- Australian National University
- Bio-Lab, University of Genoa, Italy
- Council for the Central Laboratory of the Research Councils (CCLRC), UK
- CC-IN2P3, France
- Distributed Framework, Singapore
- Distributed Aircraft Maintenance Environment (DAME), UK
- eMinerals Project, UK
- eScience, Belfast Center
- Fraunhofer ITWM, Germany
- High Energy Accelerator Organization, KEK, Japan
- K* Grid Computing, Korea
- KEK Computing Center, Japan
- Lyon, France
- NorGrid, Norway
- Nanyang Data Grid, Singapore
- Queensland University of Technology (QUT), Australia
- Rutherford Appleton Laboratory (RAL), UK
- T-Systems, Germany
- UK eScience Project, UK
- UniGrid, Poland
- UMK, Poland
- Virtual Laboratory for eScience, Netherlands

Problem

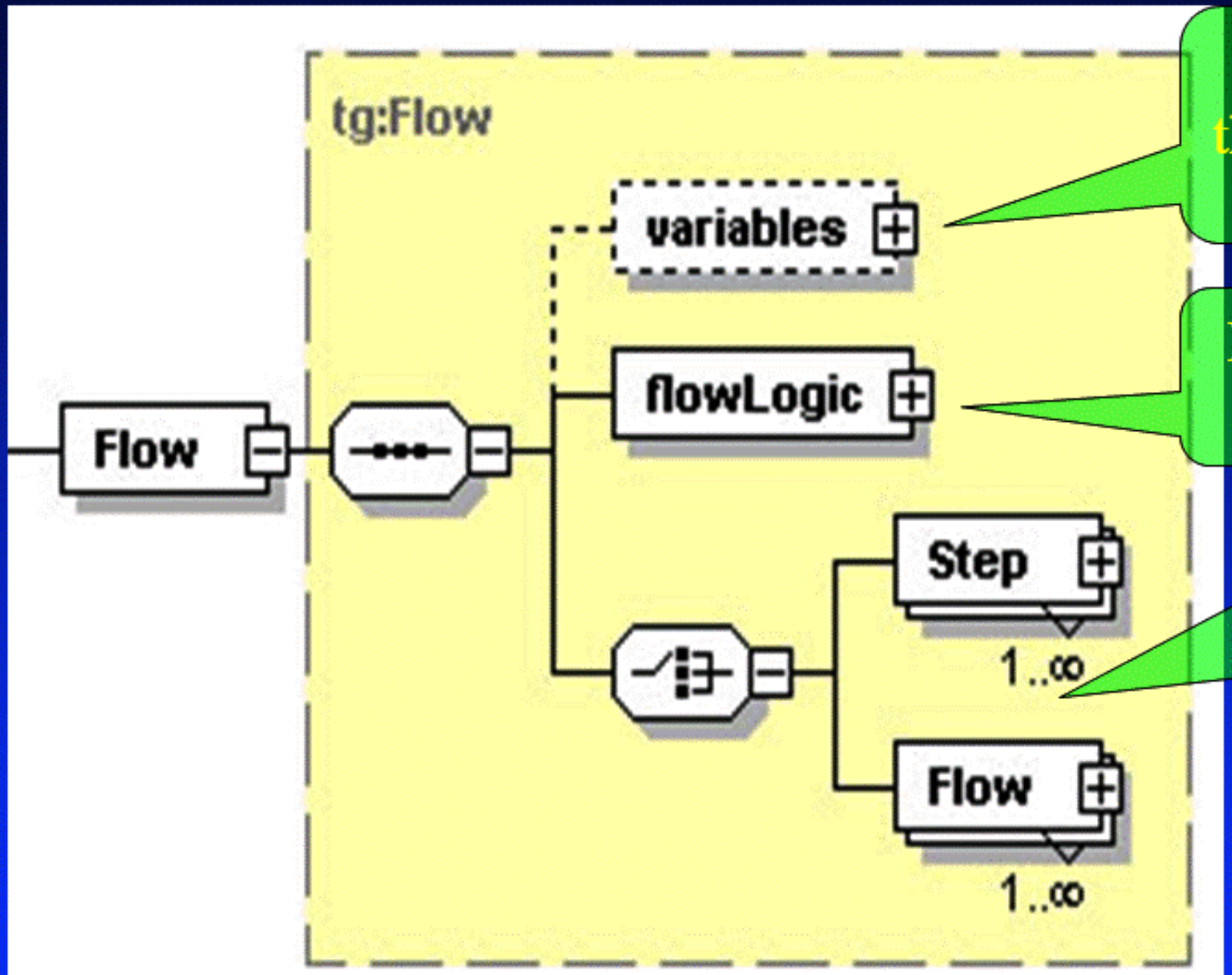
- **In short, we got rebels and misfits (in a nice way)**
  - Any file system event (like insert) on the Data Grid should start a long-run process on the file system
- **Data Storage Workflow Automation Language**
  - Need a way to describe the workflow (currently scripts)
  - Need more flexible, dynamic ways, to *describe, manage and query data storage workflow*

Solution

Data Grid Language (DGL)

- **XML based gridflow description**
  - Describes data storage execution flow logic
  - Implemented as a web service
- **ECA-based rule description for execution**
  - ECA = Event, Condition, Action
- **Querying of Status of Gridflow**
  - XQuery / Simple query of a Gridflow Execution
- **Scoped variables and gridflow patterns**
  - For, For-each, parallel, sequential, switch-cases

Flow

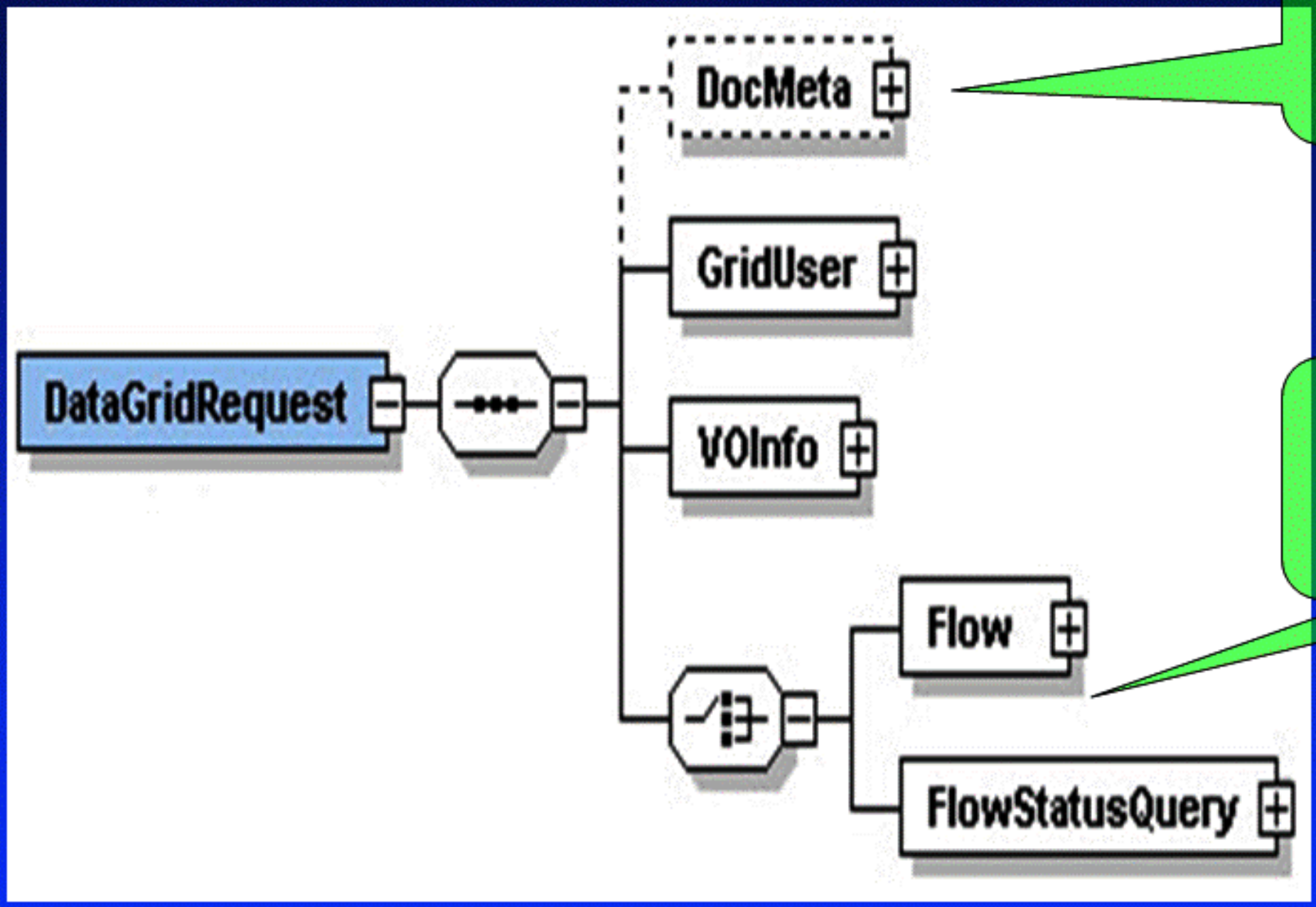


Scoped Variables that can control the flow

Logic used by the sub-members

Sub-members that are the real execution statements

Data Grid Request



Annotations about the Data Grid Request

Can be either a Flow or a Status Query

Concepts

- ECA : Event-Condition-ActionS
- Infrastructure Independent Process Description on Infrastructure independent logical storage
- Data Storage related operations and variables in workflow language
- Dynamic workflow description update
- Data Grid Workflow patterns

Status

Just graduated from our “time well wasted” lab…

From prototypes now to production software, next steps…

- Open source collaborative development project
- UCSD/SDSC, UCSB, UK CCLRC, Australia University, commercial company (-ies)
- Its very easy to write a driver and take advantage of our infrastructure

Acknowledgment

- SDSC Data Grid Group (SDSC SRB Software)
- SDSC Storage (Infrastructure and Production)
- SDSC Matrix Team