Distributed Computing for LIGO Data Analysis

The Aspen Winter Conference on Gravitational Waves, (GWADW) at Isola d'Elba, Italia
22 May 2002

Albert Lazzarini
LIGO Laboratory
Caltech

LIGO-G020232-00-E

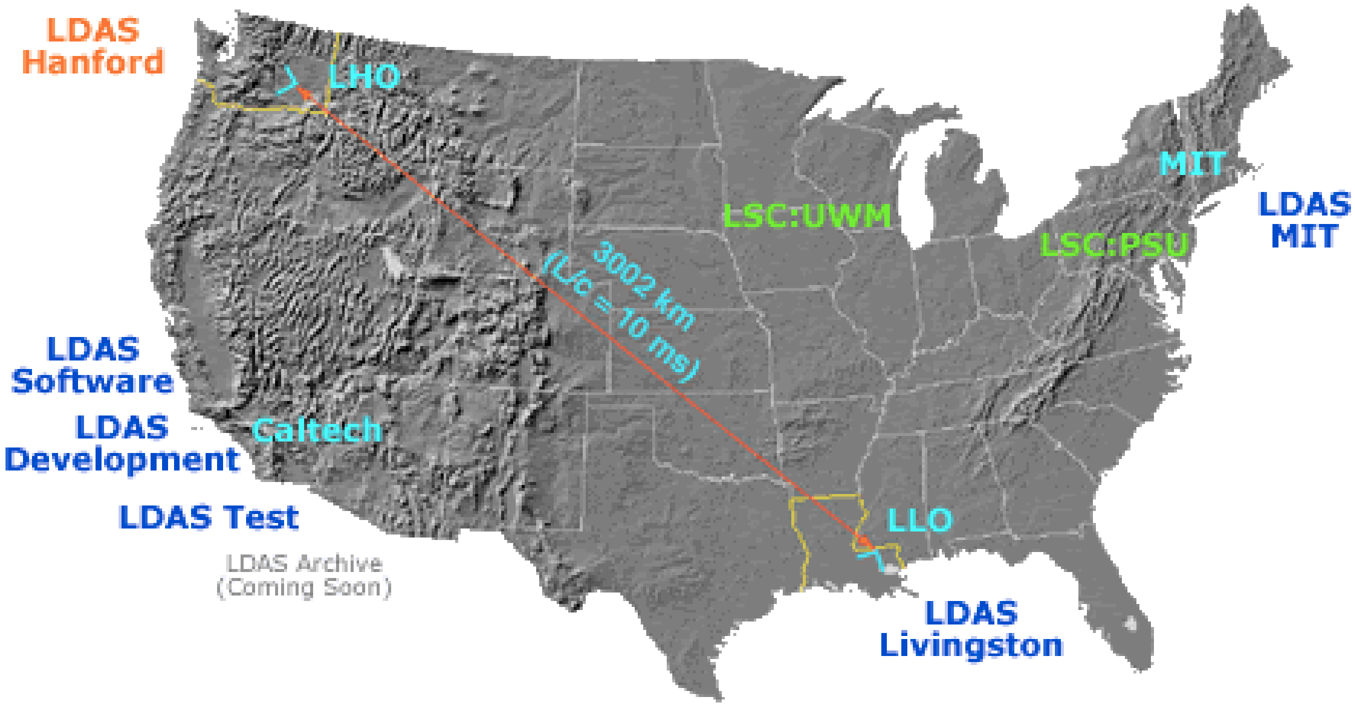
LIGO Laboratory Data Analysis System (LDAS)

A distributed network of resources within LIGO Laboratory and its Collaboration

(http://www.ldas-sw.ligo.caltech.edu)

Geographically Dispersed Laboratory plus Collaboration Institutional Facilities

Distributed Computing Has Been Necessarily Part of the LIGO Design from the Beginning



LIGO is in the GriPhyN Collaboration

GriPhyN = Science Applications + CS + Grids

- GriPhyN = Grid Physics Network (NSF Program)

US-CMS	High Energy Physics
US-ATLAS	High Energy Physics
LIGO/LSC	***Gravitational wave research***
SDSS	Sloan Digital Sky Survey

  - Strong partnership with computer scientists

- Design and implement *production-scale* grids
  - Develop common infrastructure, tools and services
  - Integration into the 4 experiments
  - Application to other sciences via “Virtual Data Toolkit” (VDT)

- Multi-year project
  - ***GriPhyN - R&D for grid architecture : 5 years, starting 2000***
  - ***iVDGL - implementation of initial Tier 2 Centers for LIGO: 5 years, starting 2001***
    - Integrate Grid infrastructure into experiments through VDT middleware software

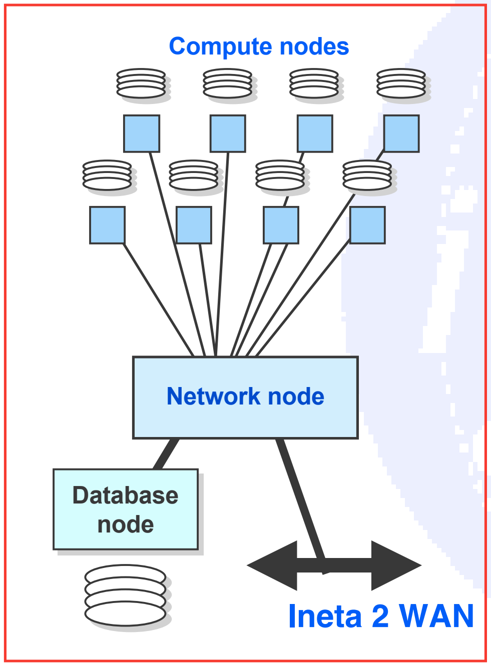
Institutions working on LIGO grid research

- *LIGO Laboratory*
  - *Tier 1 Center - GriPhyN, iVDGL*
    - ***Caltech** -- main archive, data center*
    - ***MIT** - laboratory-operated Tier 2 Center*
    - ***Observatories** - data generation centers*
- LIGO Scientific Collaboration (LSC)
  - Tier 2 Centers
    - **University of Wisconsin at Milwaukee** - GriPhyN, iVDGL
    - **Pennsylvania State University** - iVDGL
  - Tier 3 Centers & outreach
    - **University of Texas at Brownsville** (Hispanic minorities) - GriPhyN, iVDGL
    - **Salish-Kootenai College, Montana** (Native American tribal college) - iVDGL
- *Computer Science*
  - *University of Southern California/ISI (Kesselman et al.)*

Tiered Grid Hierarchical Model for LIGO

(Grid Physics Network Project - http://www.griphyn.org)

Grid Tier 2 Node:
N compute + Database + Network

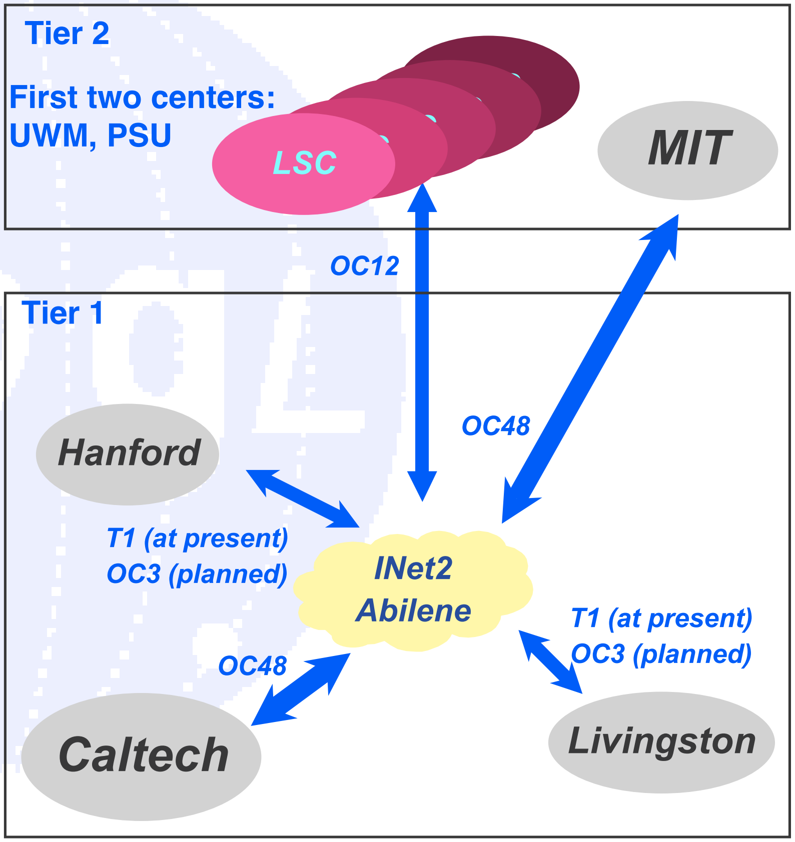


Compute nodes

Network node

Database
node

Ineta 2 WAN



Tier 2

First two centers:
UWM, PSU

LSC

MIT

OC12

OC48

Tier 1

Hanford

T1 (at present)
OC3 (planned)

INet2
Abilene

T1 (at present)
OC3 (planned)

OC48

Caltech

Livingston

LIGO and LSC Computing Resources Serve Multiple Uses

		Updated 2002.03.01	*Resource Usage Model for LSC Computing*										
			LIGO Laboratory							LSC Institutions			Other Grid Collaborators
		Function	DMT	CIT-Dev (LDAS)	CIT-Test (LDAS)	CIT-Production (LDAS)	LHO (LDAS)	LLO (LDAS)	MIT (LDAS)	PSU Tier II, iVDGL	UWM Tier II, iVDGL	UTB Tier III	USC/ISI
Scientific & infrastructure Software Development	1	***LDAS Software Development***		🟥 Priority 1 Color	🟦 Priority 2 Color				🟩 Priority 3 Color				
	2	***LDAS Integration & Tests***		🟥	🟥					🟦	🟦	🟦	
	3	LDAS CVS Software Distribution		🟥 Primary Site	🟩 Available Mirror Site	🟩 Available Mirror Site	🟩 Available Mirror Site	🟩 Available Mirror Site	🟩 Available Mirror Site				
	4	***LAL Software Development***							🟥	🟥	🟥	🟥	
	5	***LAL Scientific Validation***			🟦				🟥	🟥	🟥	🟥	
	6	***LAL integration & Test Validation***		🟥	🟦				🟥	🟥	🟥	🟥	
	7	LAL CVS Software Distribution		🟦 Sencondary Mirror Site						🟦 Sencondary Mirror Site	🟥 Primary Site	🟩 Available Mirror Site	
Data Archival & Reducttion	8	Production: Level 1 Data					🟥	🟥					
	9	Archive/Distribute Level 1 Data				🟥							
	10	Production: Level 2 Data				🟥	🟦	🟦					
	11	Archive/Distribute Level 2 Data				🟥			🟥 Subset of Level 2	🟥 Subset of Level 2	🟥 Subset of Level 2		
	12	***Production: Level 3 Data***				🟥	🟦	🟦		🟥	🟥		
	13	***Archive/Distribute Level 3 Data***				🟥	🟥	🟥	🟥	🟥	🟥		
Data Analysis	14	***On-site Searches***					🟥	🟥	🟥				
	15	***Off-site Searches***				🟥	🟩	🟩	🟥	🟥	🟥	🟥	
	16	***Multiple Detector Analysis***				🟥			🟥	🟥	🟥	🟥	
	17	Monte Carlo Runs				🟦			🟥	🟥	🟥	🟥	
Grid R&D	18	***Detector Characterization***	🟥						🟥	🟥	🟥	🟥	
	19	***Grid SW Development***		🟩	🟩					🟥	🟥	🟥	🟥
	20	***Grid SW Integration & Testing***		🟩	🟩					🟥	🟥	🟥	🟥
	21	Numerical GR & Source Simulations											
	22	Hardware Simulations	General Computing Resources within LIGO										

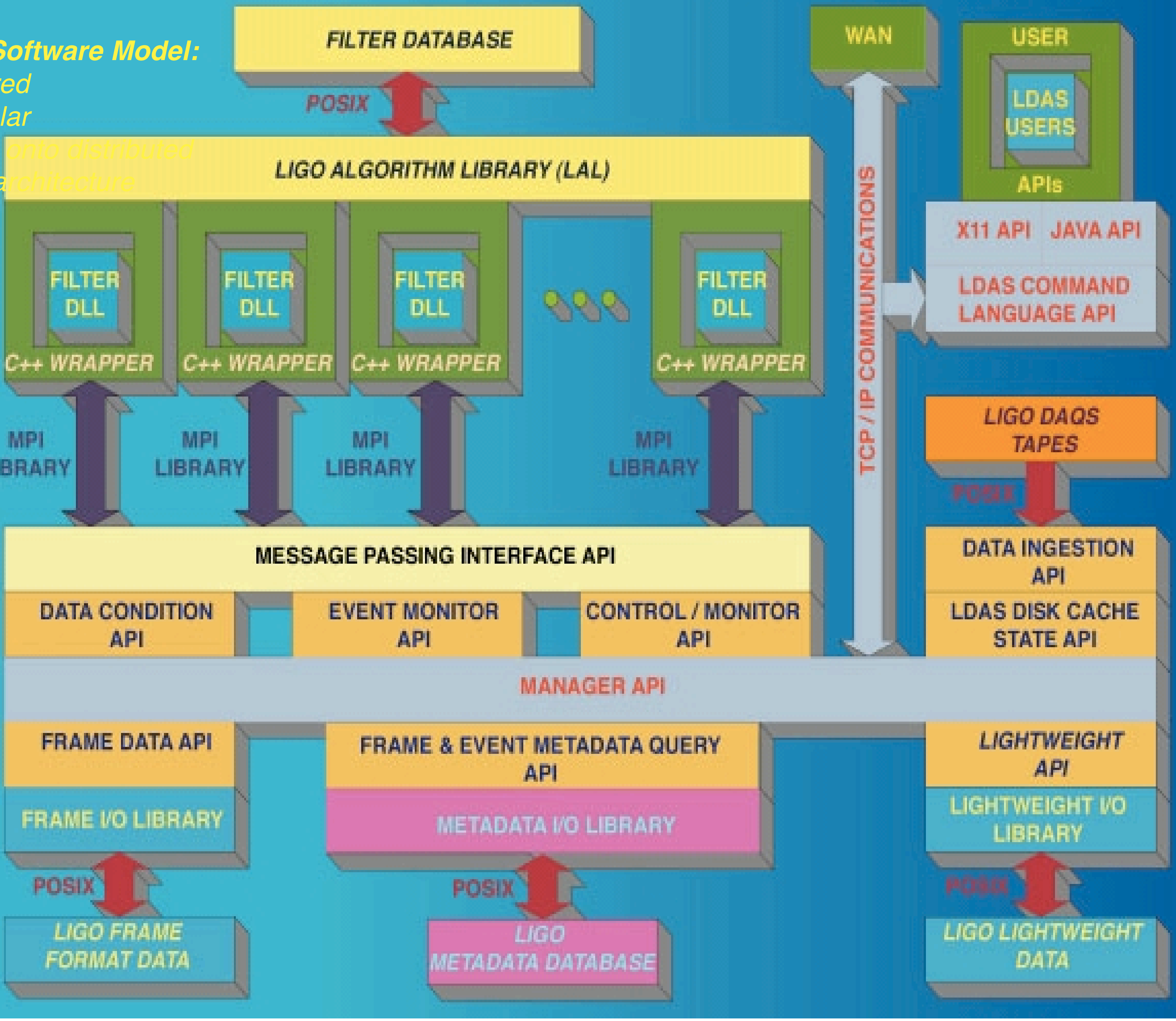
Priority Legend

Priority 1	🟥
Priority 2	🟦
Priority 3	🟩

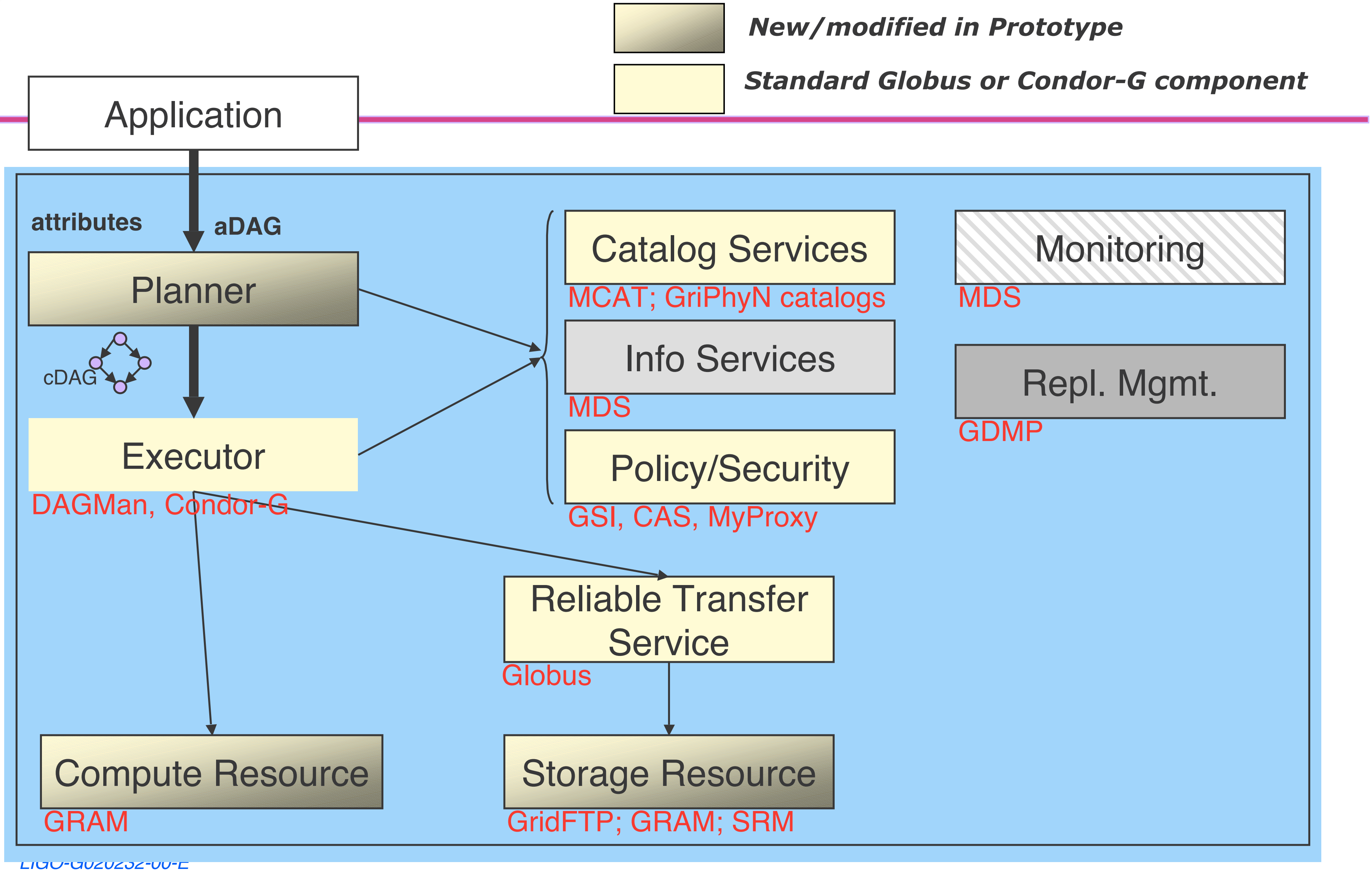
http://www.ldas-sw.ligo.caltech.edu

LIGO Software Model:

- *Layered*
- *Modular*
- *Maps onto distributed H/W architecture*



Preliminary GriPhyN Data Grid Architecture



LIGO data & processing needs that can be fulfilled by the grid

- data replication -

- **LIGO archive replica**
  - » 40TB today
  - » 300TB by 2003-2004
  - » Transposed, reduced data sets archived remotely efficient access by collaboration users from second source
    - – Tier 2 centers
    - – Teragrid (Caltech/SDSC/ANL/NCSA)
  - » Geographic separation from Tier 1 center at Caltech
    - – Redundant access
    - – Faster access for other U.S. regions

LIGO data & processing needs that can be fulfilled by the grid

- extended computational resources on the grid -

- **Massively parallel processing of GW channel**
  - **Inspiral searches to low mass**
    - e.g.: ***[5-50 Mflop/byte]*** for inspiral search of GW channel
    - ***x [0.2 TB]*** total cleaned GW channel for LIGO I
    - Science analysis software maintained by Collaboration as a vetted, valdiated body of scientific software
      - LAL -- LIGO Algorithm Library
      - Dynamically loaded libraries (DLLs), shared objects (DSOs)
      - Loaded at run time per script specification from CVS archive
  - **Large-area search for unknown periodic sources**
    - Long (coherent) Fourier Transforms
      - For weakest signals
      - e.g., 1 kHz for 10 days => ~ 10^9 point FFTs
    - Barycenter motion modulates signal uniquely for every point in the sky
    - *The CW equivalent of the "filter bank"*

➛Unlike other grid projects, LIGO has data *NOW*

- ***Strategic use of US national computing resources to extend LIGO and Collaboration capabilities***

LIGO-G020232-00-E

Grid research within LIGO

CY2001 - CY2002

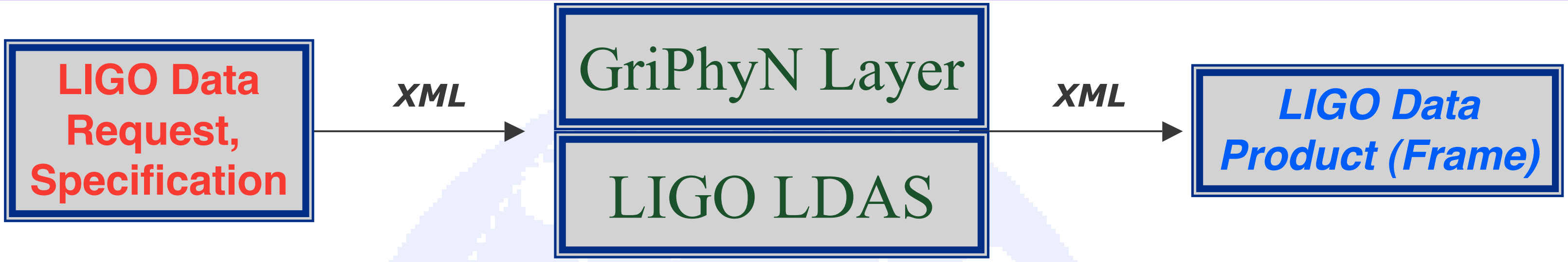
- Developed LIGO Virtual Data requirements
- LIGO/GriPhyN prototype

  Simple demonstration of Virtual Data Concepts (SuperComputing 2001 Convention)

  - Data access transparency with respect to location
  - Data access transparency with respect to materialization
- Provided a Globus interface to LDAS
  - Basis for a secure access to LIGO resources
- Designed the Transformation Catalog
  - Needed to find an appropriate executable binaries for a given H/W architecture
  - Can be used in many systems
- Basic infrastructure for the development of Virtual Data concepts
  - Foundation for Year 2

GriPhyN/LIGO

prototype functionality



LIGO Data Request, Specification

XML

GriPhyN Layer

LIGO LDAS

XML

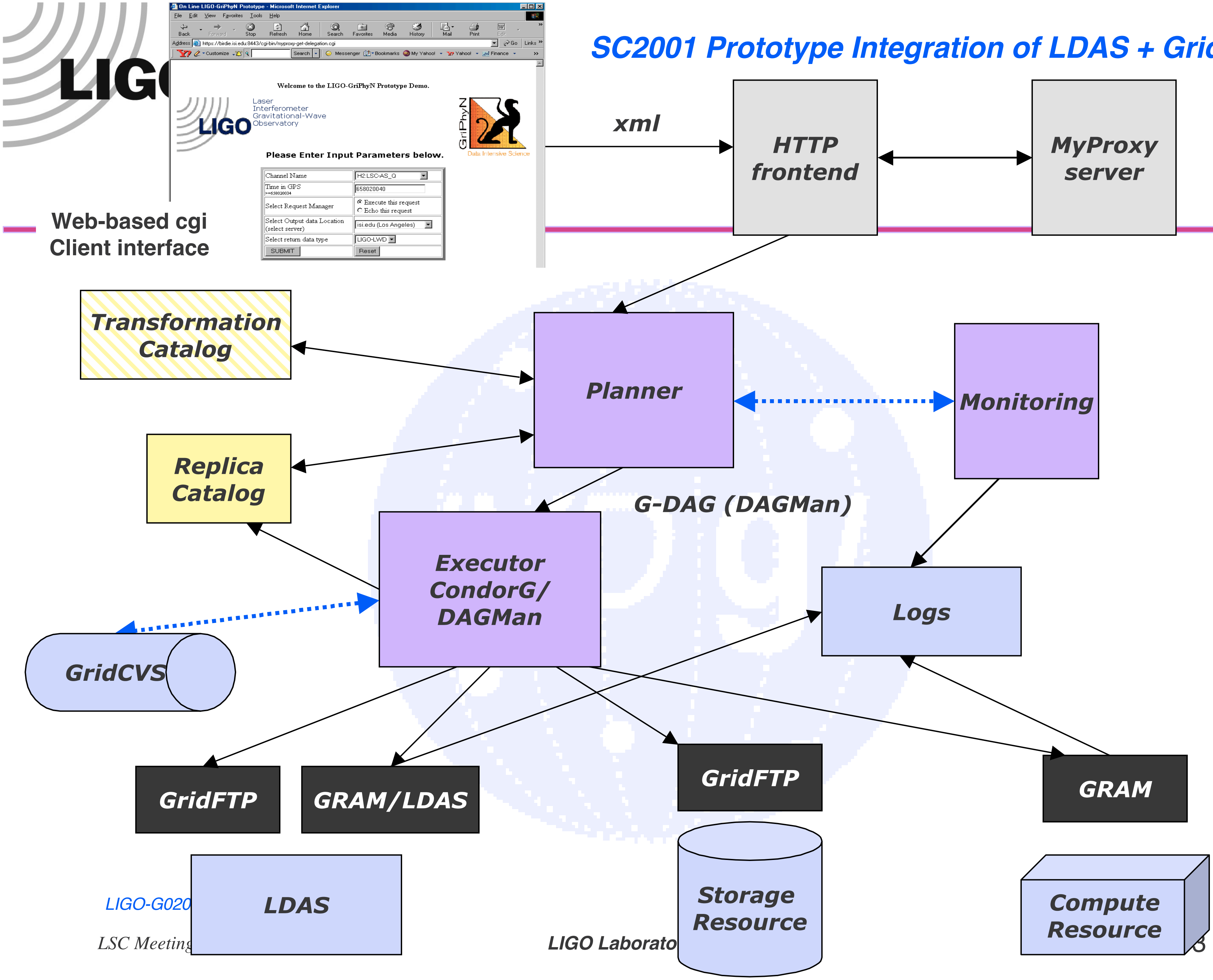
LIGO Data Product (Frame)

- Interpret an XML-specified request
- Acquire user’s proxy credentials
- Consult replica catalog to find available data
- Construct a plan to produce data not available
- Execute the plan
- Return requested data in Frame or XML format

Year 1 Virtual Data Product: single channel frame: Extraction

Compute resources running LIGO Data Analysis System at Caltech and UWM, storage resources at ISI, UWM and Caltech

SC2001 Prototype Integration of LDAS + Grid Tools



On Line LIGO-GriPhyN Prototype - Microsoft Internet Explorer

Address: https://birdie.isi.edu:8443/cgi-bin/myproxy-get-delegation.cgi

Welcome to the LIGO-GriPhyN Prototype Demo.

LIGO Laser Interferometer Gravitational-Wave Observatory

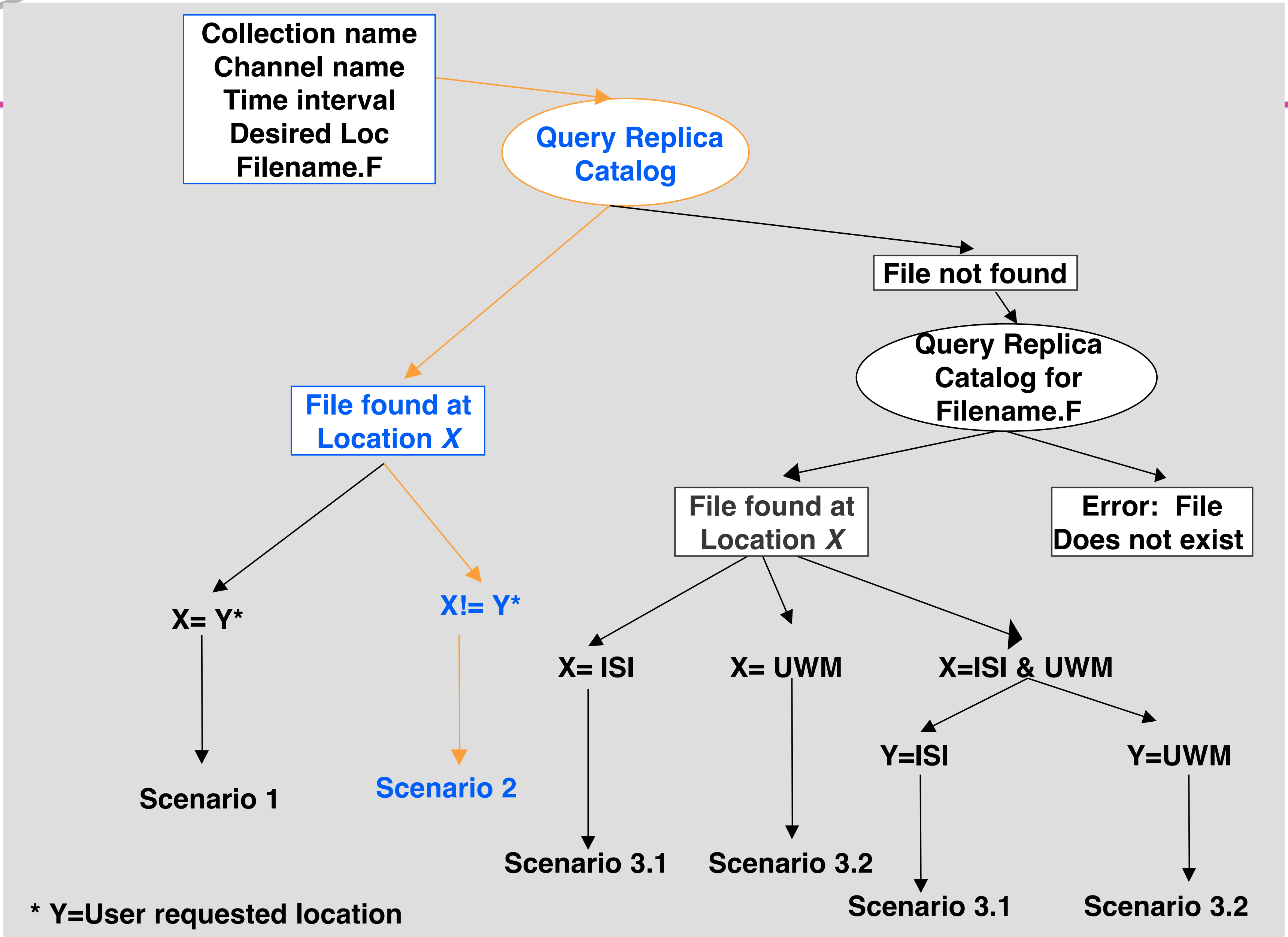
GriPhyN Data Intensive Science

Please Enter Input Parameters below.

Channel Name	H2:LSC-AS_Q
Time in GPS >=658020034	658020040
Select Request Manager	● Execute this request ○ Echo this request
Select Output data Location (select server)	isi.edu (Los Angeles)
Select return data type	LIGO-LWD
SUBMIT	Reset

LIGO GriPhyN

Planner for virtual data requests



Collection name
Channel name
Time interval
Desired Loc
Filename.F

Query Replica Catalog

File not found

Query Replica Catalog for Filename.F

File found at Location *X*

X= Y*

X!= Y*

Scenario 1

Scenario 2

File found at Location *X*

Error: File Does not exist

X= ISI

X= UWM

X=ISI & UWM

Y=ISI

Y=UWM

Scenario 3.1

Scenario 3.2

Scenario 3.1

Scenario 3.2

* Y=User requested location

Pulsar Search Mock Data Challenge

Extending CY2001 Prototype

- Extend prototype beyond data access and requests
  - Large-area GW pulsar search, as a science focus
  - Use of virtual data (“SFTs”)
  - Request planning and execution of analysis on distributed resources
- Broaden the GRAM/LDAS interface
  - Richer variability and functionality in data access methods:
    - Short time Fourier transforms as virtual data (SFTs)
    - Concatenation, decimation and resampling or frame data
- Design a Data Discovery mechanism for discovery of data replicas.
  - Ability to interact with the LDAS Diskcache resources
- Implementation of the Data Discovery mechanism to support the pulsar search

Year 2 Research & Development

- Explore bulk data operations
  - Finding new available data
  - Registering data into catalogs
- Deepen the understanding of Virtual Data naming
  - How do you ask for what you want?
- Planning and Fault Tolerance
  - Need to specify model
  - Explore existing planning solutions
  - Examine fault tolerance issues at the system level
- Scalable pulsar search to scientifically interesting levels of sensitivity at SC’2002

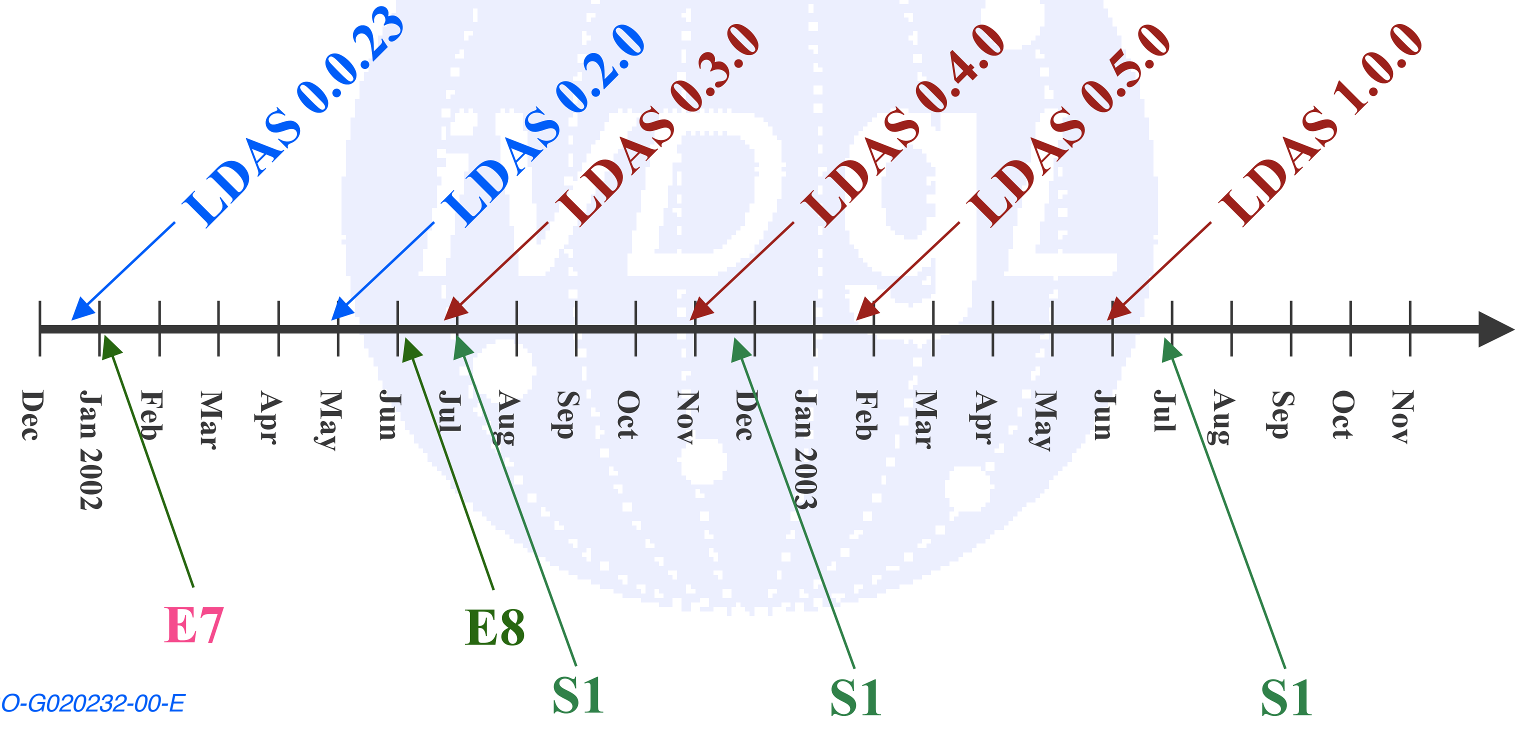
THE CHALLENGE FOR LIGO:

and working!!!!

Integrating grid functionality ***within an existing***^framework

LIGO Software is already on a production schedule!

LDAS Release Timeline



LDAS 0.0.23
LDAS 0.2.0
LDAS 0.3.0
LDAS 0.4.0
LDAS 0.5.0
LDAS 1.0.0

Dec
Jan 2002
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan 2003
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov

E7
E8
S1
S1
S1

A “Modest Proposal”

For inter-project development of grid infrastructure

1. ***NDAS*** is a working set of unix-level interfaces that effectively interleave the data of 5 different international efforts ***TODAY***
2. Use the existing and growing interaction with ***NDAS*** as a first step to developing a GW international grid (EUGrid + iVDGL)
   - ***“BREAK”*** NDAS ***temporarily*** in order to migrate the infrastructure to grid-based utilities and tools:
     - Globus package for
       - Secure, authorized data access and transmission
       - Robust data transmission using the Gridftp protocol
     - Condor-G for (***eventually***) submitting analysis jobs across collaborations
3. Add people to the NDAS team who know and can implement these technologies

FINIS