LIGO as a Data Intensive User of Teragrid

Teragrid Applications Workshop

Argonne National Laboratory

30 April 2002

Albert Lazzarini

LIGO-G020225-00-E

LIGO Laboratory Data Analysis System (LDAS)

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

Geographically Dispersed Laboratory plus Collaboration Institutional Facilities

LIGO-G020225-00-E

LIGO and LSC Computing Resources Serve Multiple Uses

Scientific & infrastructure Software Development

Data Archival & Reducttion

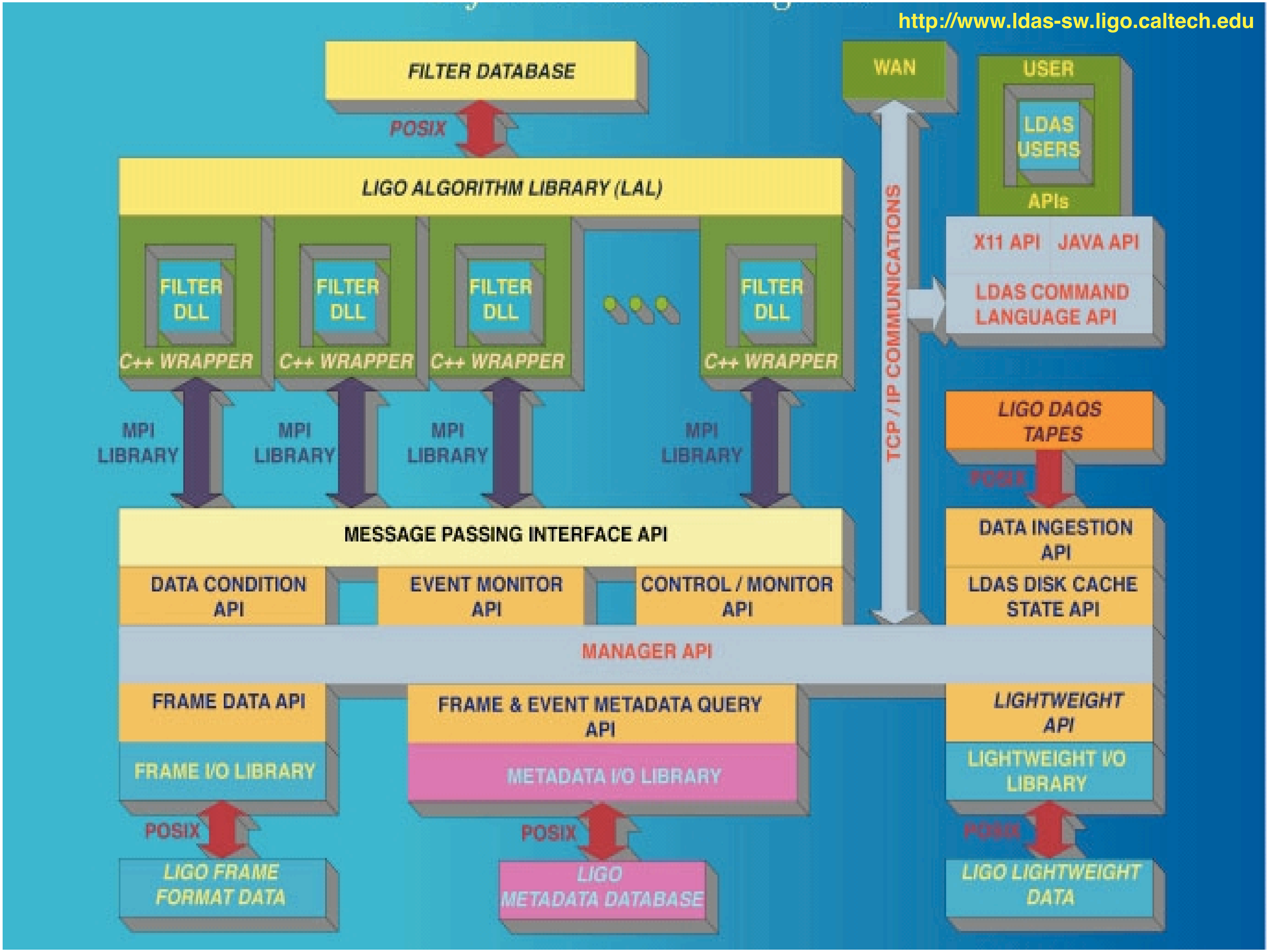
Data Analysis

Grid R&D

	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
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	
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***											
14	***On-site Searches***											
15	***Off-site Searches***											
16	***Multiple Detector Analysis***											
17	Monte Carlo Runs											
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
FILTER DATABASE
POSIX
WAN
USER
LDAS USERS
APIs
LIGO ALGORITHM LIBRARY (LAL)
FILTER DLL
C++ WRAPPER
FILTER DLL
C++ WRAPPER
FILTER DLL
C++ WRAPPER
FILTER DLL
C++ WRAPPER
TCP / IP COMMUNICATIONS
X11 API
JAVA API
LDAS COMMAND LANGUAGE API
MPI LIBRARY
MPI LIBRARY
MPI LIBRARY
MPI LIBRARY
LIGO DAQS TAPES
POSIX
MESSAGE PASSING INTERFACE API
DATA INGESTION API
DATA CONDITION API
EVENT MONITOR API
CONTROL / MONITOR API
LDAS DISK CACHE STATE API
MANAGER API
FRAME DATA API
FRAME & EVENT METADATA QUERY API
LIGHTWEIGHT API
FRAME I/O LIBRARY
METADATA I/O LIBRARY
LIGHTWEIGHT I/O LIBRARY
POSIX
POSIX
POSIX
LIGO FRAME FORMAT DATA
LIGO METADATA DATABASE
LIGO LIGHTWEIGHT DATA

LIGO Data Analysis

- Continuous Time series
  - » 16 kHz, 160 Hz, 1 Hz....
- Analysis performed in both time and frequency domains
  - » Single channel over long period of time - source searches
  - » Many channels over short period of short time - regression, data reduction
  - » Optimal matched Wiener filtering
    - – *90% CPU time spent in Fast Fourier Transforms*

LIGO Data & Processing needs

- **LIGO archive replica**
  - » 40TB today
  - » 300TB by 2003-2004
  - » Transposed data archived remotely on Teragrid for efficient access by collaboration users from second source
  - » Geographic separation from Tier 1 center at Caltech desirable
- **Repeated digital signal processing of GW channel**
  - » e.g.: [5-50 Mflop/byte] for inspiral search of GW channel
  - » x [0.2 TB] total cleaned GW channel for LIGO I
  - » System-based pipelines in LDAS environment
  - » Sandard scripts, interfaces
  - » Personal filters (individual exploratory research with data, LIGO Algorithm Library)
  - » DLL, DSOs loaded at run time per script specification from CVS

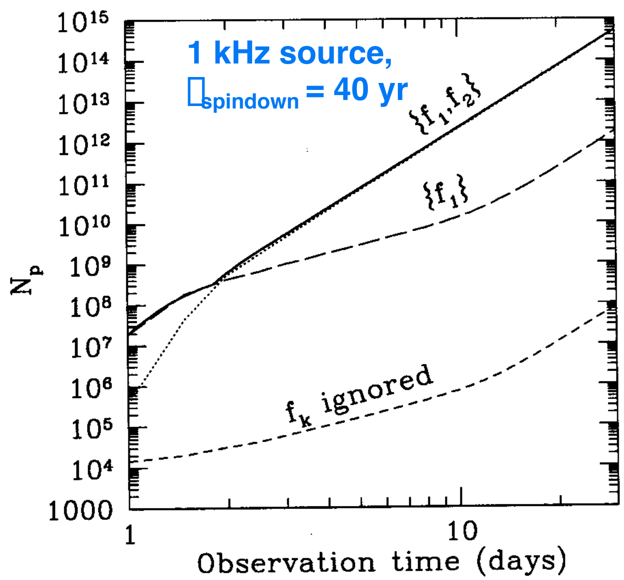
LIGO data processing challenge

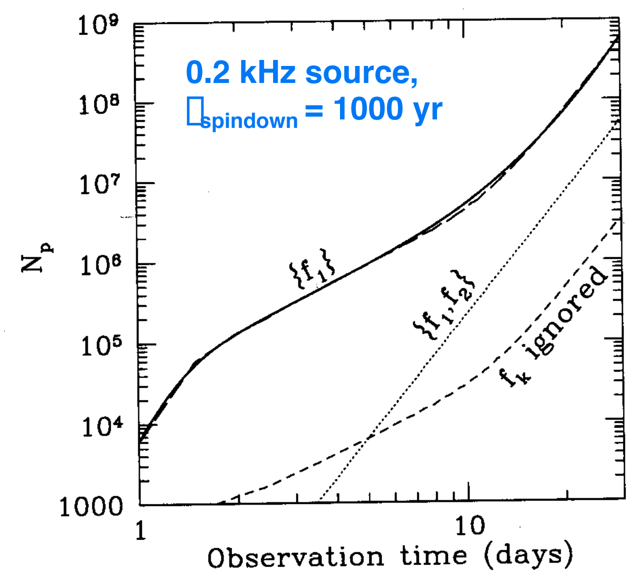
- **Large-area search for unknown periodic sources**
  - » Very long Fourier Transforms due to expected weakness of signals
    - – e.g., 1 kHz for 10 days => ~ 10^9 point FFTs
  - » Earth motion modulates signal uniquely for eavry point in the sky

THE CHALLENGE

Generally the phase evolution of the source is not known and one must perform searches over some parameter space volume

- The number of templates grows dramatically with the coherent integration time baseline and the computational requirements become prohibitive:



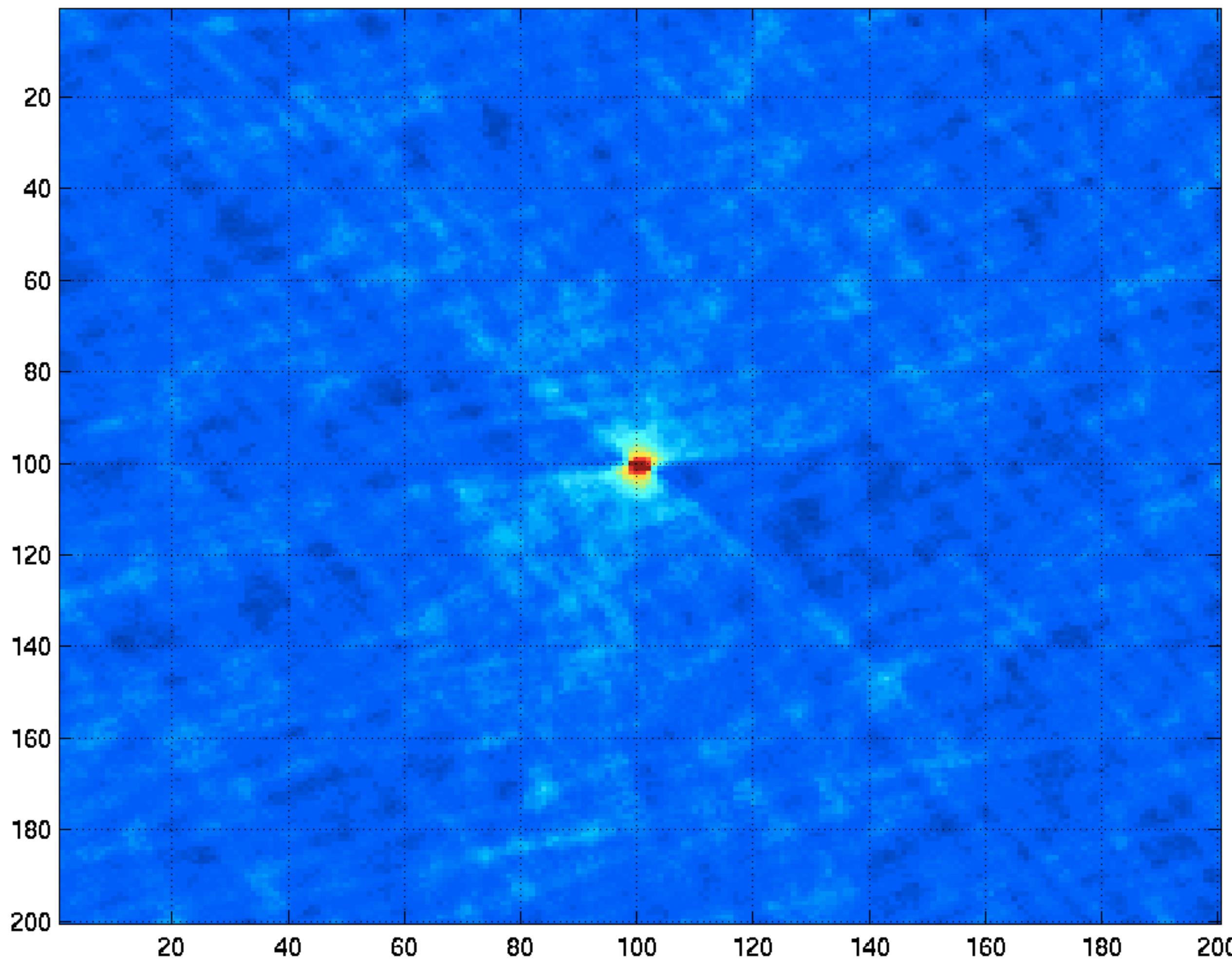


On a 1TFLOPS computer it would take more than 10000 yr to perform an all-sky search over 1000 Hz for an observation time of 4 months.

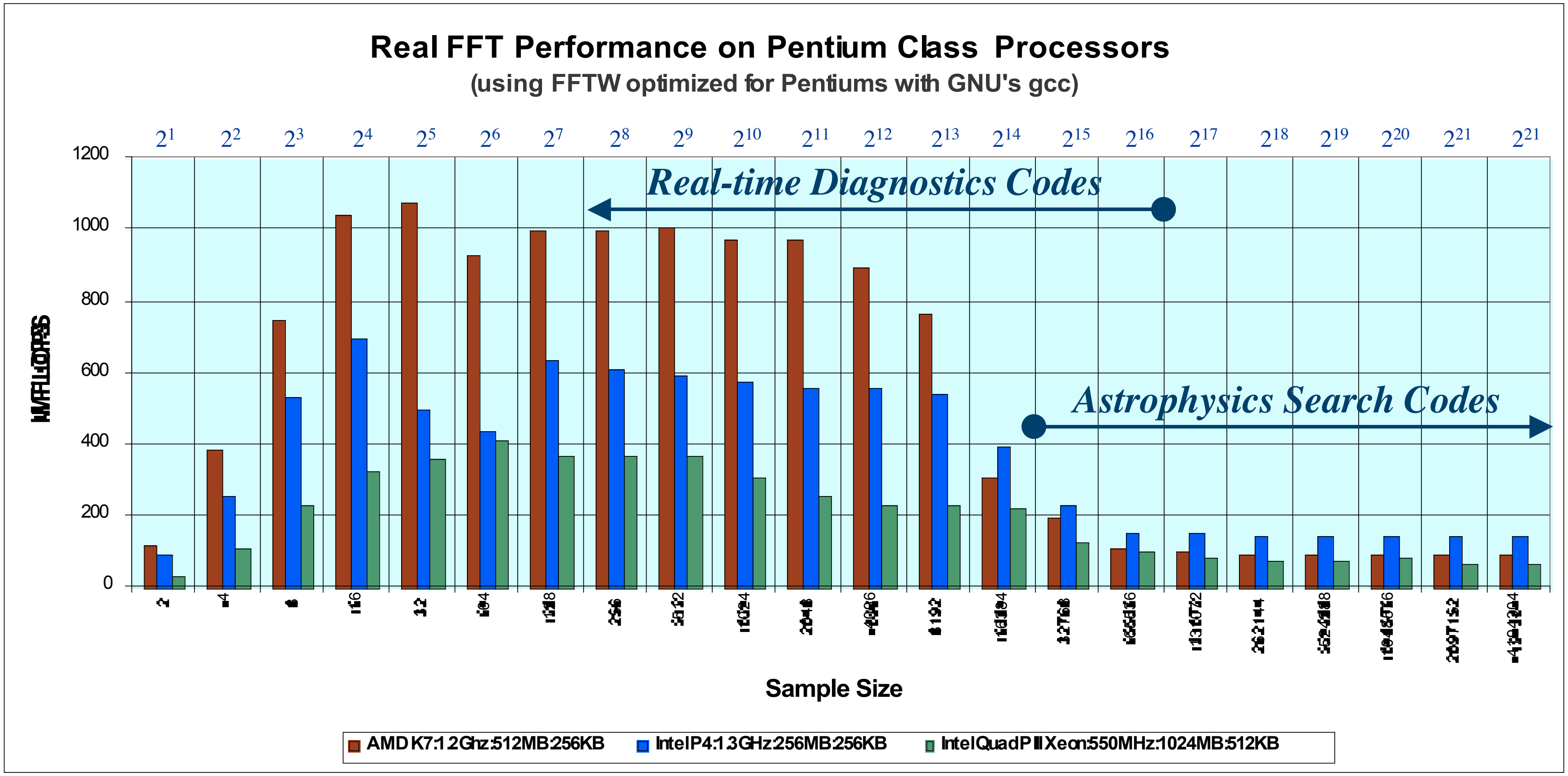
** Graphs from Brady, Creighton, Cutler, and Schutz, gr-qc/9702050*

Simulated Hough Transform Image

- **Image:**
  - » **8 hours of integration per DeFT(column)**
  - » **Total observation time of roughly 3 months.**
  - » **SNR is such that 129 out of the 270 signal points were registered.**
  - » **The source is located at alpha=45 delta=45 degrees.**
  - » **The source's intrinsic frequency is 400 Hz**
  - » **Signal has no spindown.**



Recent CPU Node Performance



- **Pipeline analysis of LIGO data computationally dominated by cost of Fast Fourier Transforms (FFT).**
  - *Non-Hierarchical Binary Inspiral Search spends an average of ~90% of CPU cycles performing FFT.*
- **Most practical/efficient data segment size as much as 2^{20} points for Binary Inspiral Search.**