Data Analysis Needs to Support the Search for Gravitational Waves in the U.S.

Personnel Needs in the Grid Computing Era

L.Samuel Finn[2], Erik Katsavounidis[1], Albert Lazzarini[1],
Gary Sanders[1], Alan Wiseman[3]

[1]LIGO Laboratory
[2]The Pennsylvania State University
[3]University of Wisconsin at Milwaukee (LSC)

Meeting at NSF to discuss ITR 2003
8 November 2002
Washington, D.C.

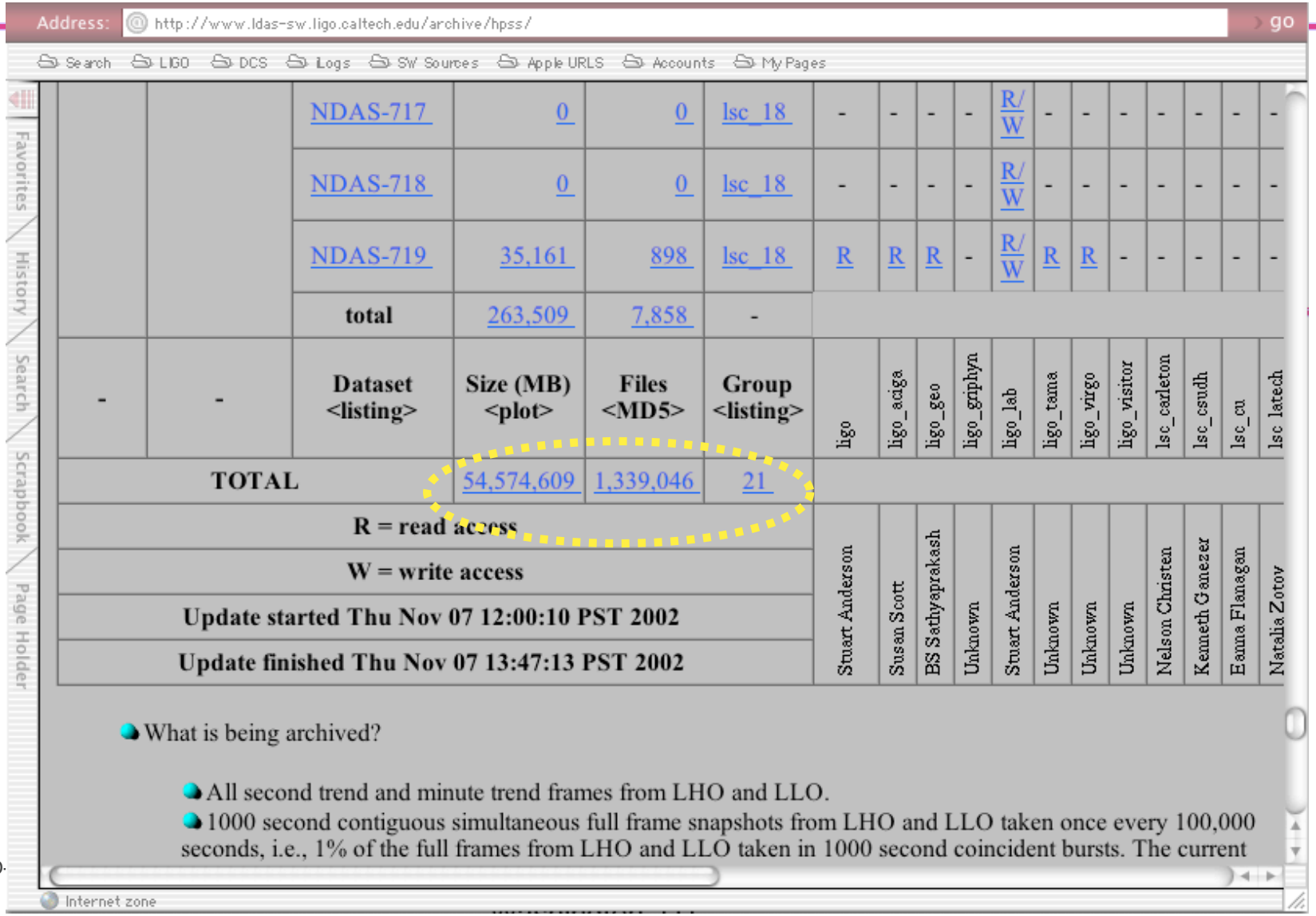
Outline

- Prologue -- LIGO science runs have started
- How did we get where we are today?
- Needs for the collaboration
  - » LIGO Laboratory
  - » Tier 2 Centers outside the Laboratory
- Issues and Questions

Prologue

- LIGO is taking data ***TODAY***!
  - » >54 TB from engineering runs E1 - E8, and first science run, S1
  - » S2 (mid-Feb 2003) will almost double data volume
  - » In steady state operations, we generate ~1TB/day

Snapshot of HPSS Archive at Caltech



Address: http://www.ldas-sw.ligo.caltech.edu/archive/hpss/ go

Search LIGO DCS Logs SW Sources Apple URLS Accounts My Pages

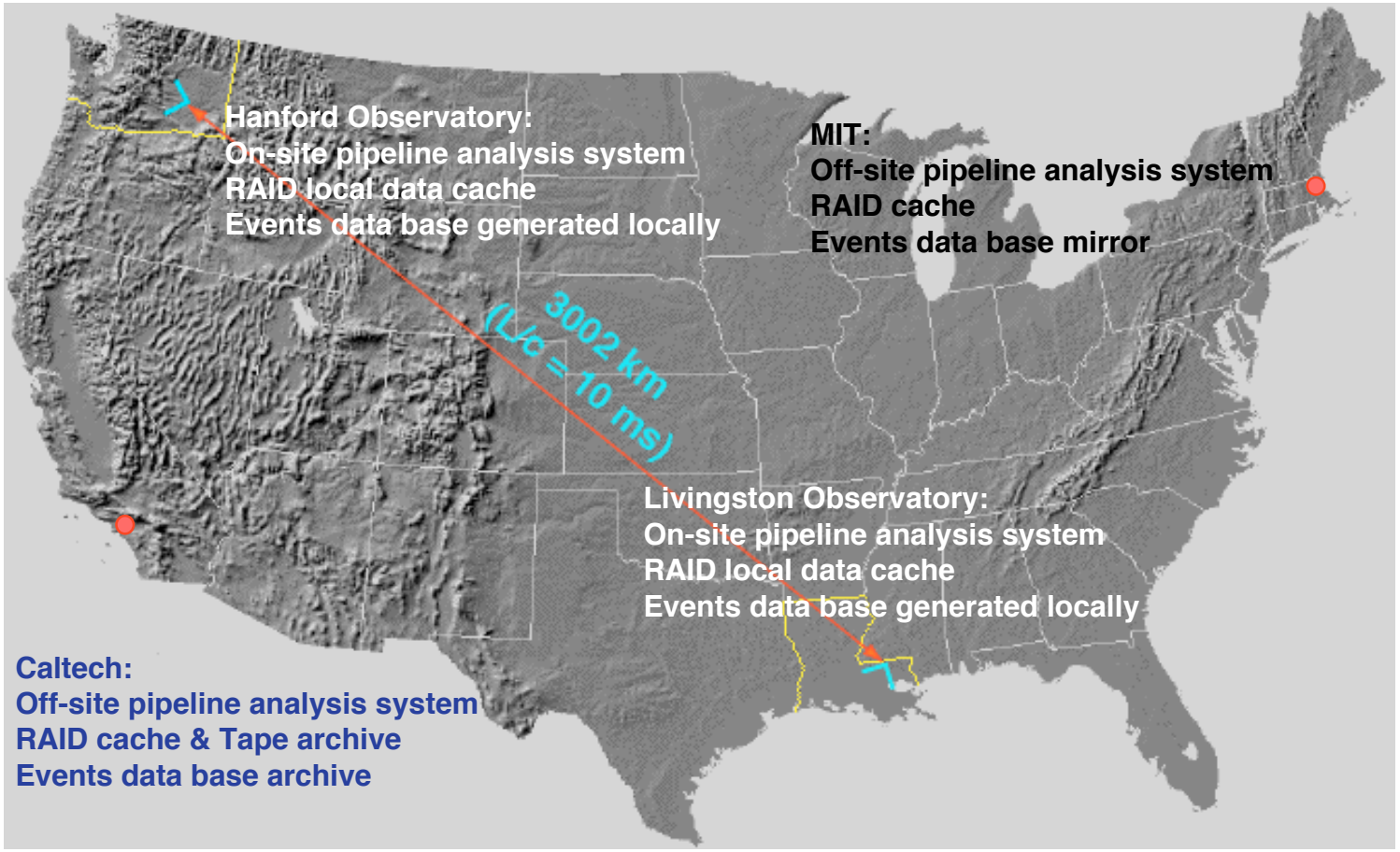
Favorites History Search Scrapbook Page Holder

		NDAS-717	0	0	lsc_18	-	-	-	-	R/W	-	-	-	-	-	-	-
		NDAS-718	0	0	lsc_18	-	-	-	-	R/W	-	-	-	-	-	-	-
		NDAS-719	35,161	898	lsc_18	R	R	R	-	R/W	R	R	-	-	-	-	-
		total	263,509	7,858	-												
-	-	Dataset <listing>	Size (MB) <plot>	Files <MD5>	Group <listing>	ligo	ligo_aciga	ligo_geo	ligo_griphyn	ligo_lab	ligo_tama	ligo_virgo	ligo_visitor	lsc_carleton	lsc_csudh	lsc_cu	lsc_latech
TOTAL			54,574,609	1,339,046	21												
R = read access						Stuart Anderson	Susan Scott	BS Sathyaprakash	Unknown	Stuart Anderson	Unknown	Unknown	Unknown	Nelson Christen	Kenneth Ganezer	Eanna Flanagan	Natalia Zotov
W = write access																	
Update started Thu Nov 07 12:00:10 PST 2002																	
Update finished Thu Nov 07 13:47:13 PST 2002																	

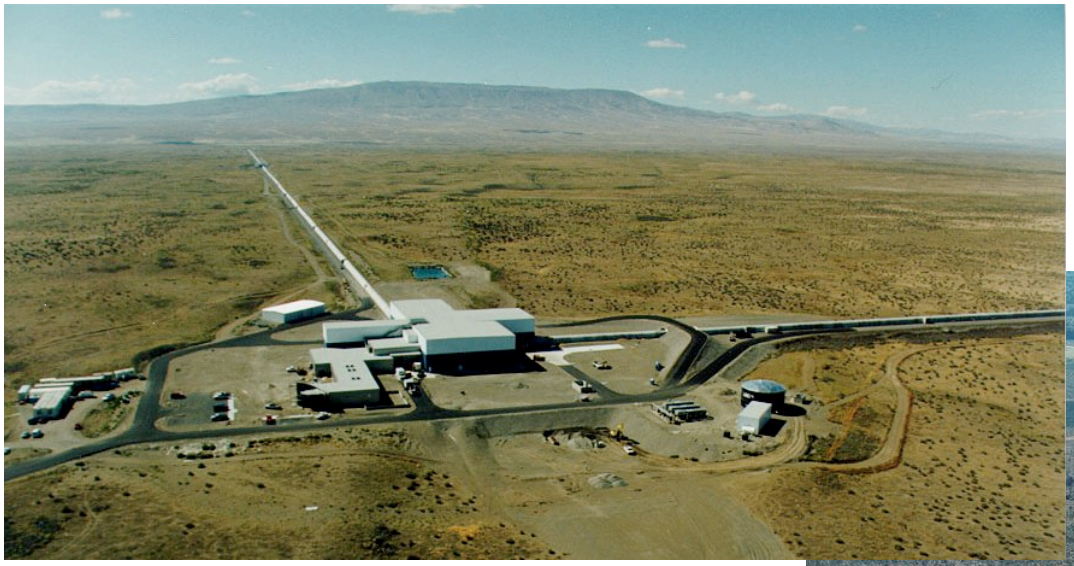
- What is being archived?
  - All second trend and minute trend frames from LHO and LLO.
  - 1000 second contiguous simultaneous full frame snapshots from LHO and LLO taken once every 100,000 seconds, i.e., 1% of the full frames from LHO and LLO taken in 1000 second coincident bursts. The current

Internet zone

The LIGO Laboratory Sites

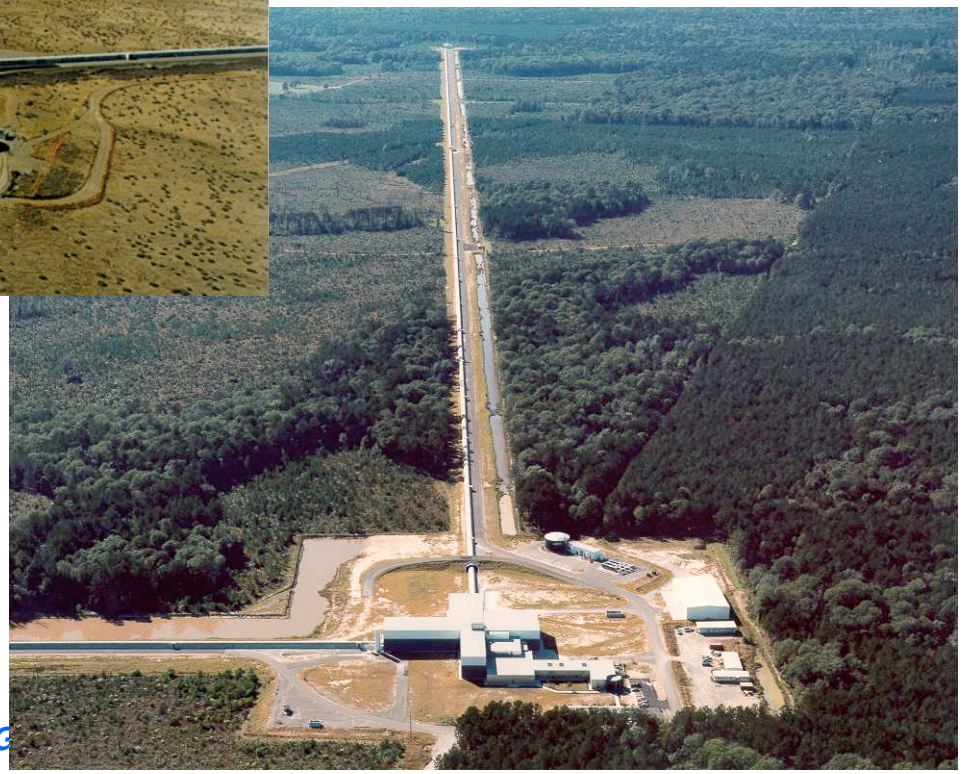


LIGO Observatories

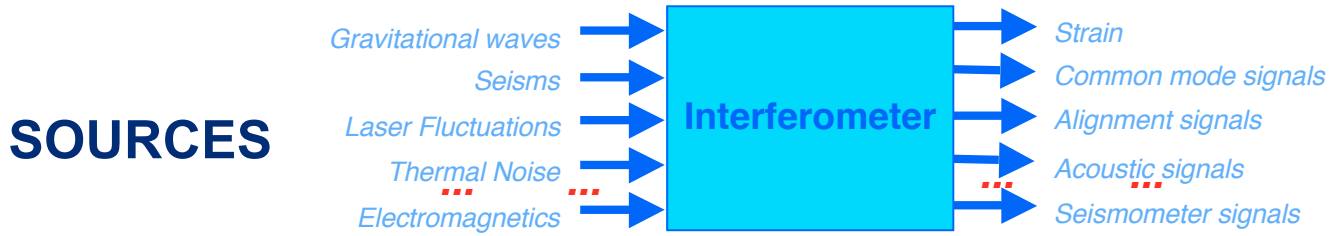


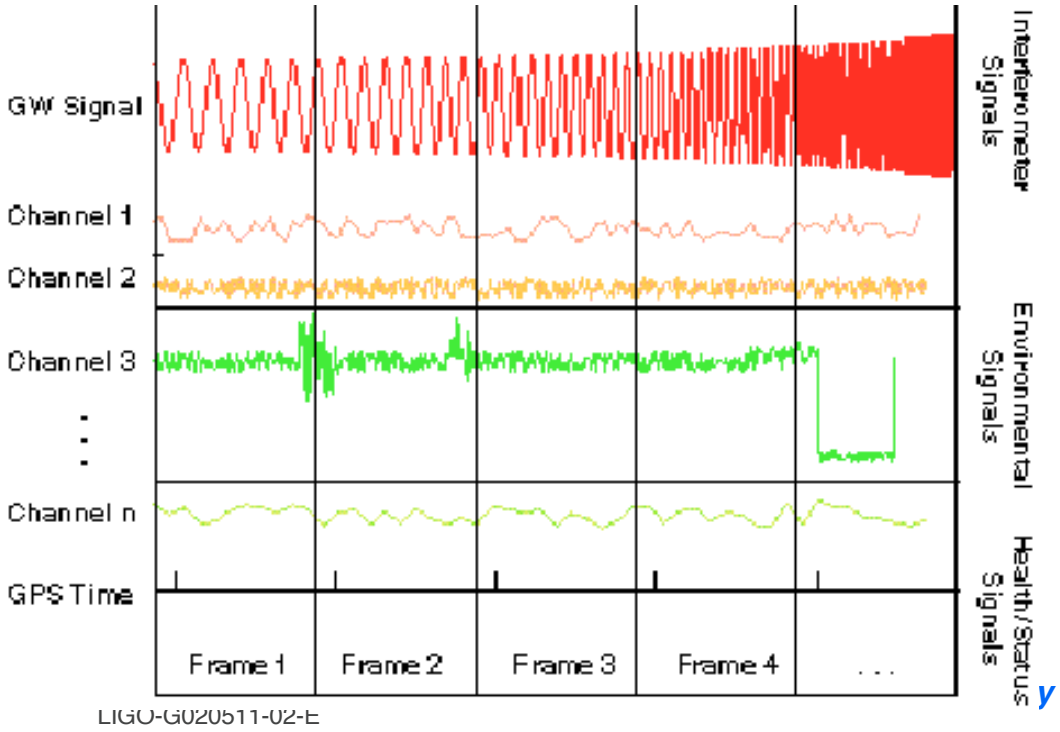
Hanford Observatory
Washington
Two interferometers
(4 km and 2 km arms)

Livingston Observatory
Louisiana
One interferometer (4km)



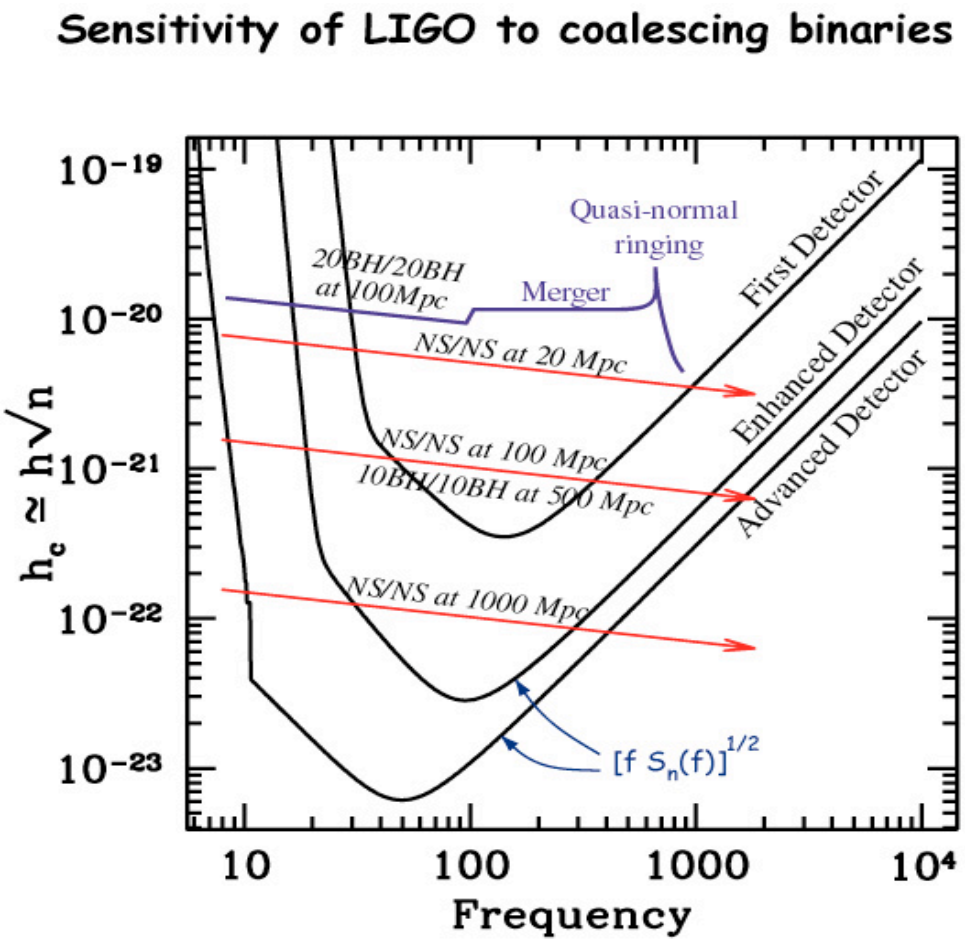
Interferometer Data

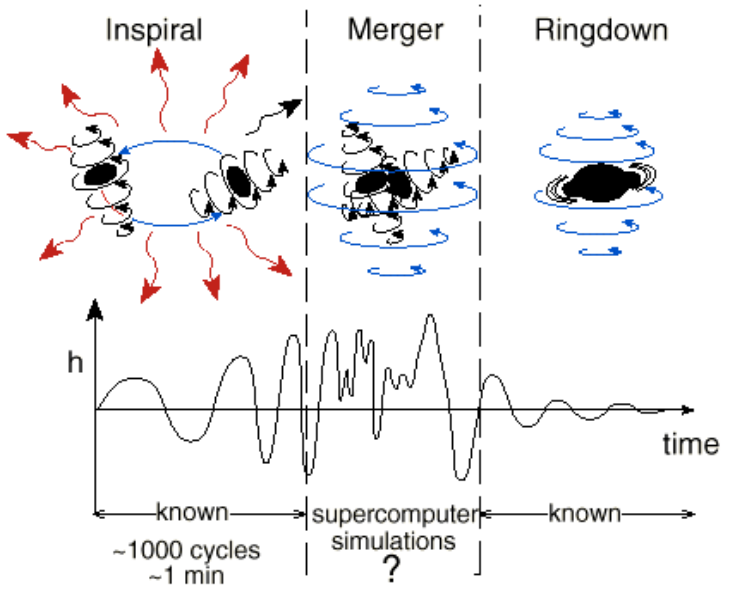




- **Frame Format: All interferometric detector projects have agreed on this standard data format.**
- **LIGO writes Frames at a rate of ~3MB/s from ~5000 channels per interferometer.**
  - ***GW Strain is ~1% of all data.***
- ***24x7 operation leads to 1TB/day for LIGO.***

Compact Binary Sources



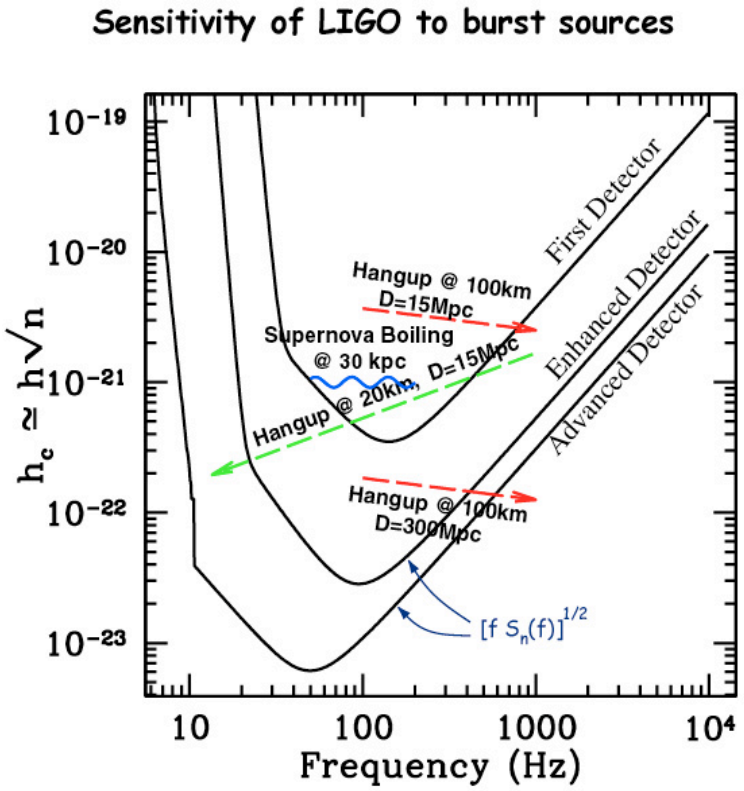


Brief Summary of Detection Capabilities of Mature LIGO Interferometers

• **Inspiral of NS/NS, NS/BH and BH/BH Binaries:** The table below [15] shows estimated rates $\mathcal{R}_{\rm gal}$ in our galaxy (with masses $\sim 1.4M_\odot$ for NS and $\sim 10M_\odot$ for BH), the distances $\mathcal{D}_{\rm I}$ and $\mathcal{D}_{\rm WB}$ to which initial IFOs and mature WB IFOs can detect them, and corresponding estimates of detection rates $\mathcal{R}_{\rm I}$ and $\mathcal{R}_{\rm WB}$; Secs. 1.1 and 1.2.

	NS/NS	NS/BH	BH/BH in field	BH/BH in globulars
$\mathcal{R}_{\rm gal}$, yr^{-1}	10^{-6}–10^{-4}	$\lesssim 10^{-7}$–10^{-4}	$\lesssim 10^{-7}$–10^{-5}	10^{-6}–10^{-5}
$D_{\rm I}$	20 Mpc	43 Mpc	100	100
$\mathcal{R}_{\rm I}$, yr^{-1}	1×10^{-4} – 0.03	$\lesssim 1 \times 10^{-4}$ – 0.3	$\lesssim 3 \times 10^{-3}$ – 0.5	0.03 – 0.5
$D_{\rm WB}$	300 Mpc	650 Mpc	$z = 0.4$	$z = 0.4$
$\mathcal{R}_{\rm WB}$, yr^{-1}	0.5 – 100	$\lesssim$ 0.5 – 1000	$\lesssim$ 10 – 2000	100 – 2000

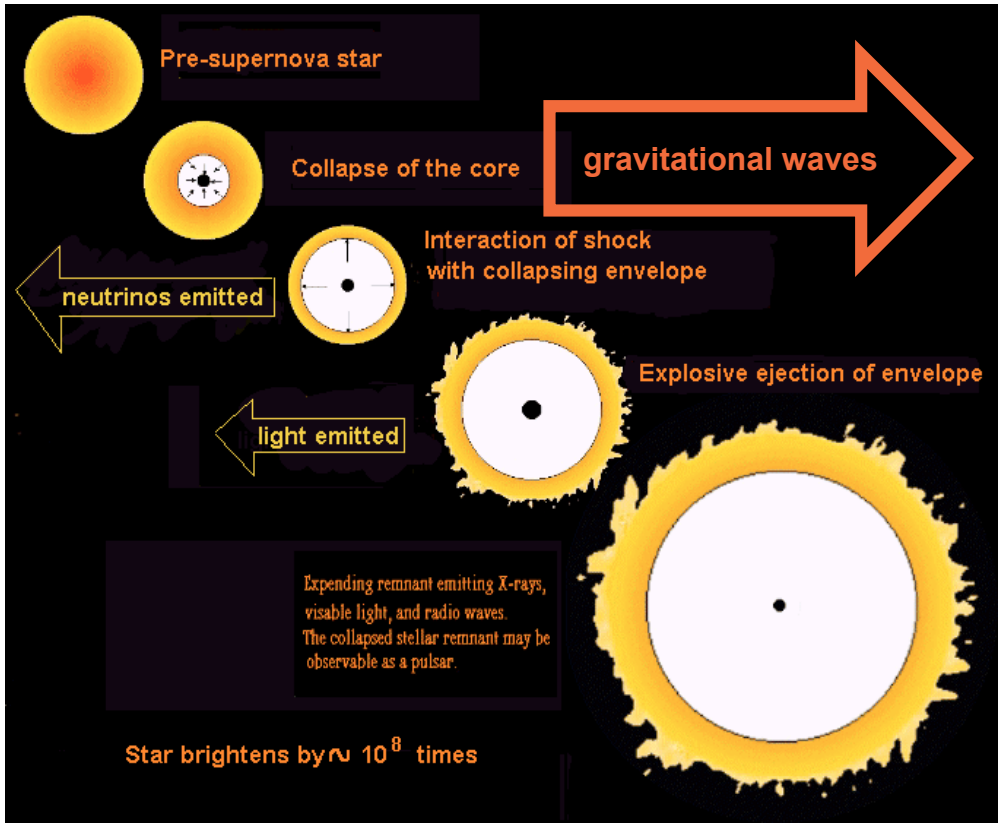
Burst Sources



Rate

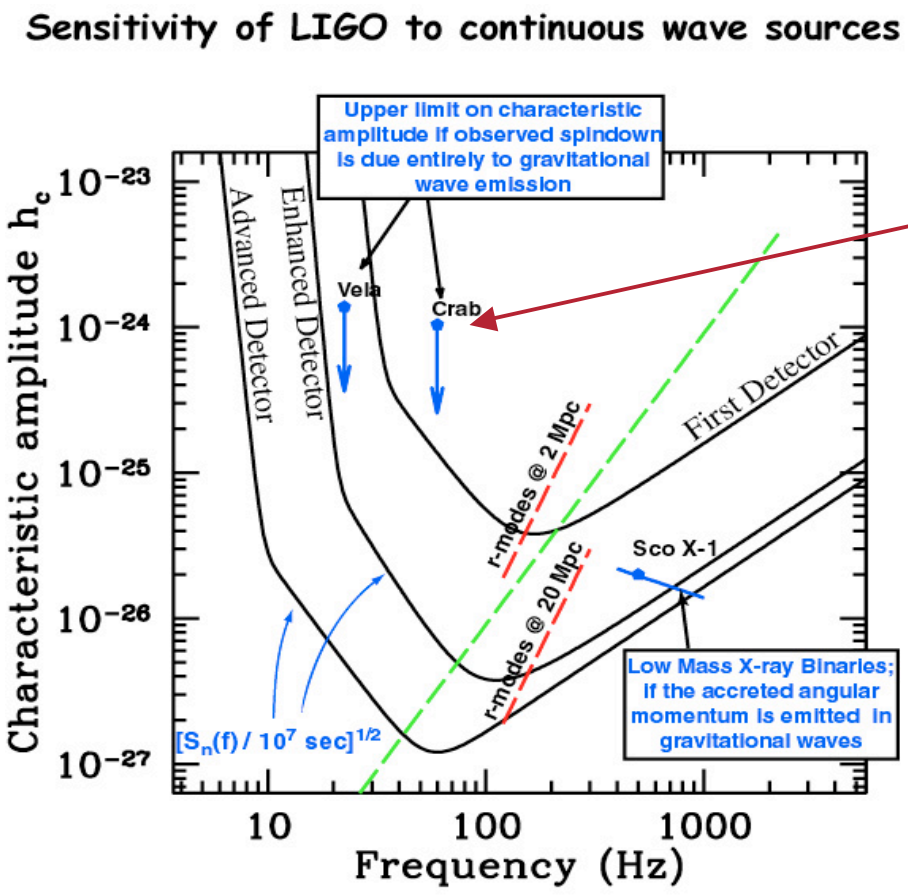
1/50 yr - our galaxy

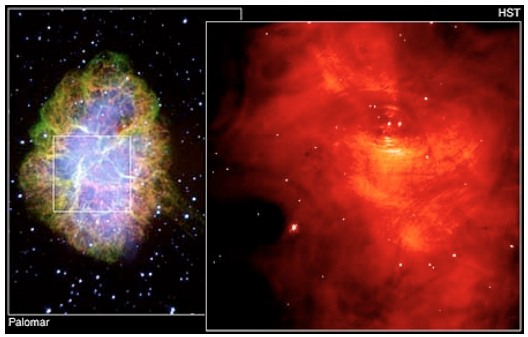
3/yr - Virgo cluster



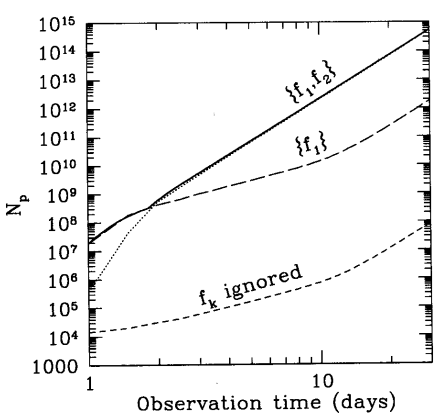
Periodic Sources

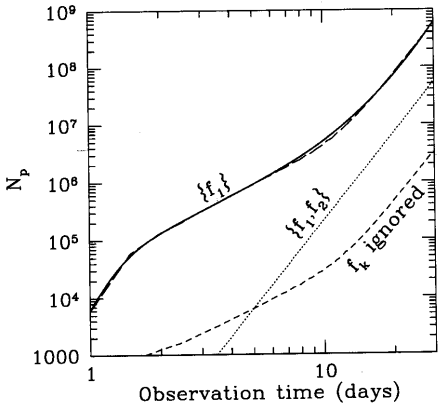
On a 1TFLOPS computer it would take more than 10,000 yr to perform an all-sky search over a 1000 Hz band for an observation time of 4 months.





Data must be corrected for each source position on the sky



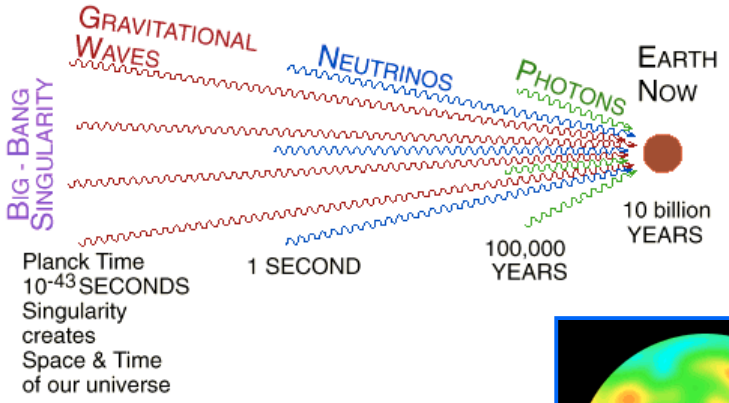


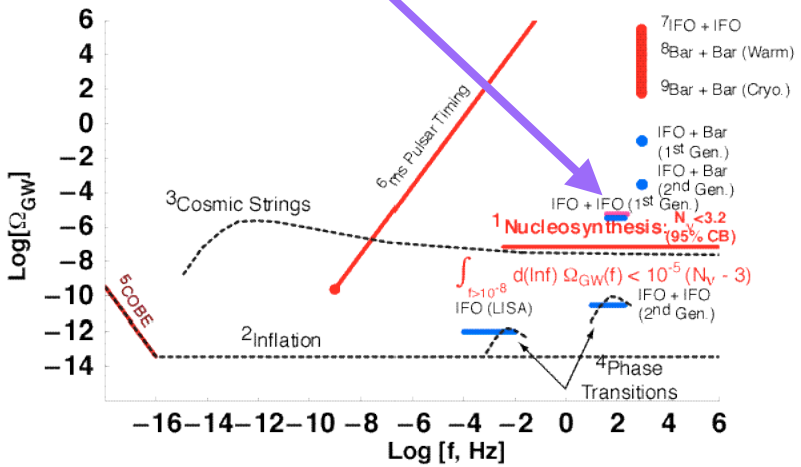
** Graphs from Brady, Creighton, Cutler, and Schutz, gr-qc/9702050*

Stochastic Background Sources

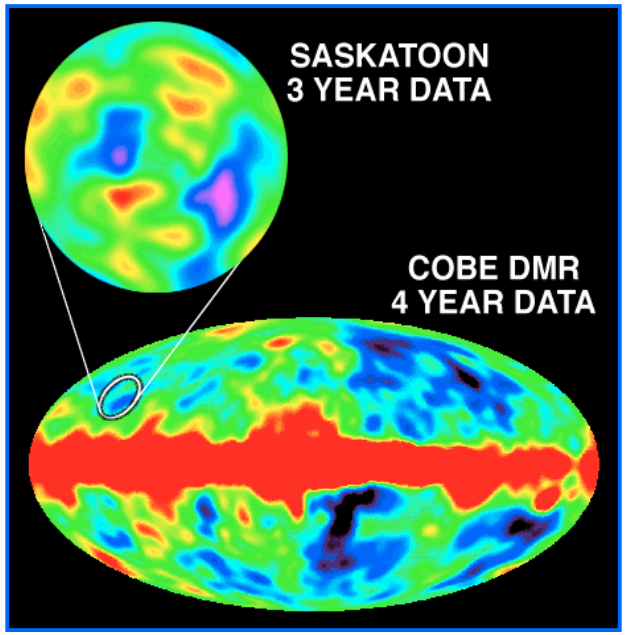
Upper limit possible with initial LIGO interferometers:

$Omega_{GW}$ ~10^{-5}





1 Kolb & Turner (The Early Universe, 1990)
Burles, Nollet,Trunan, Turner (PRL 82, 1999)
2 Grishchuk (SPJETP 40, 1975)
3 Allen & Brustein (gr-qc9609013)
Allen (gr-qc9604033)
4 Kamionkowski, Kosowoski &Turner (PRD 49, 1994)
5 Allen & Koranda (PRD 50, 1994)
6 Thorsett & Dewey (PRD 53, 1996)
Kaspi, Taylor, Ryba (ApJ 428, 1994)
7 Compton, Nicholson, Schutz, Proc. MG7 (1994)
8 Hough, Pugh, Bland, Drever, Nature 254 (1975)
9 Astone, et. al., Astr. Astroph. 351 (1999)



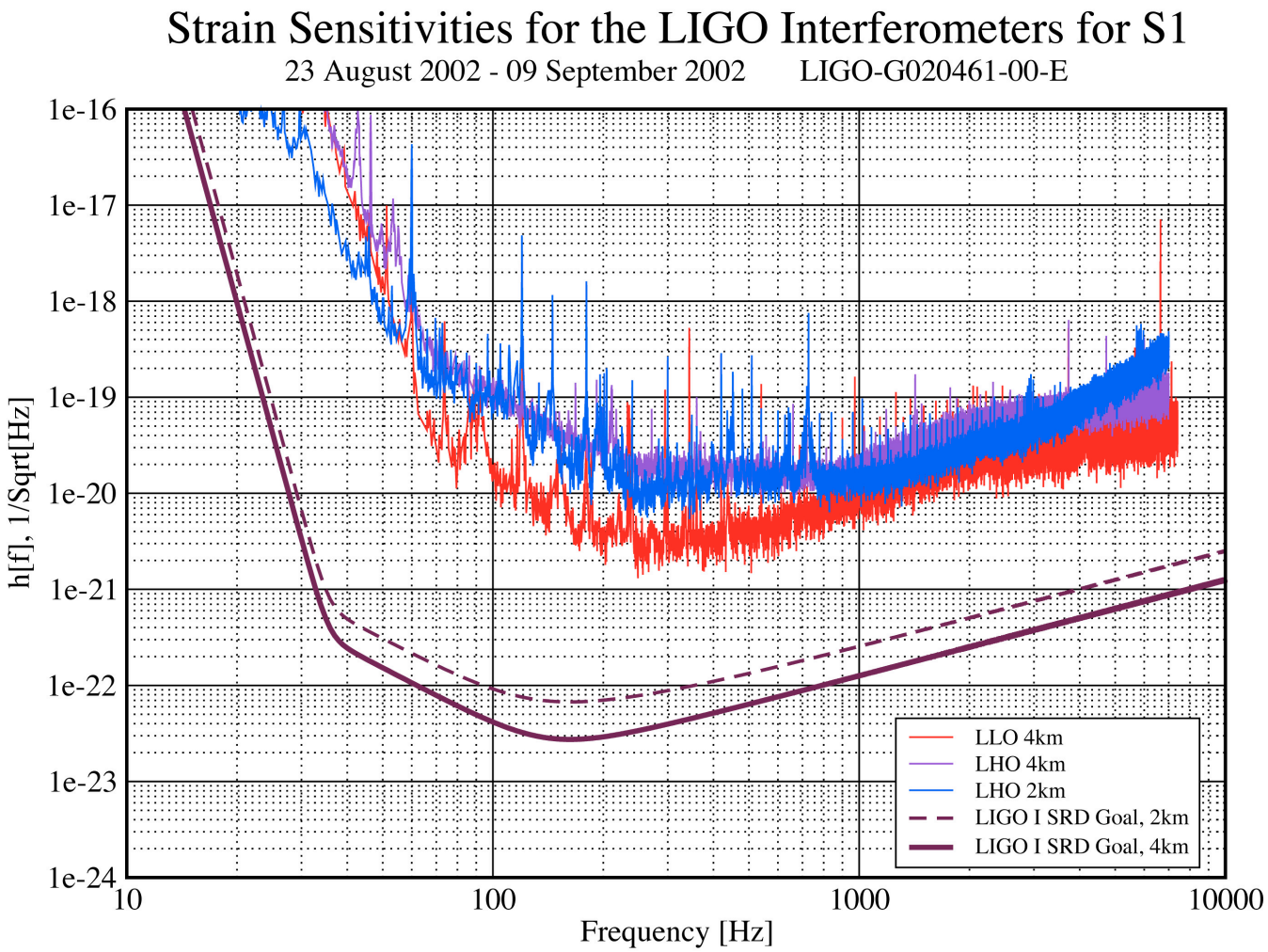
Analog from electromagnetic spectrum

LIGO Sensitivity at Start of S1

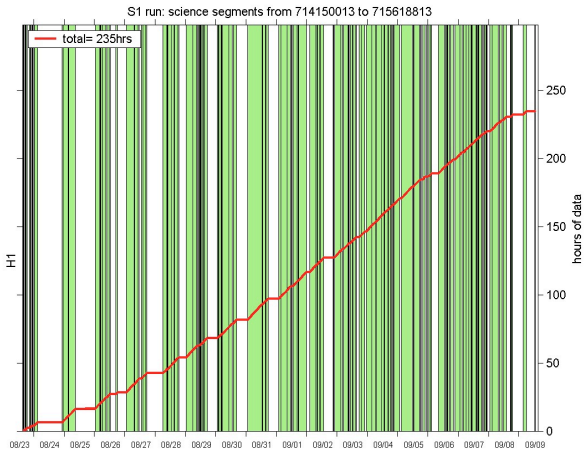
LIGO
S1 Run

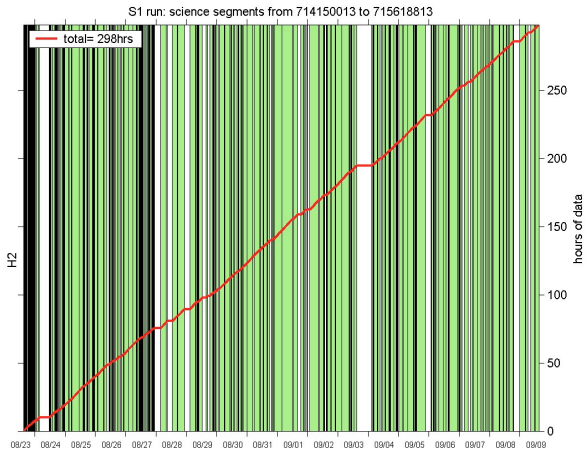
"First Upper Limit Run"

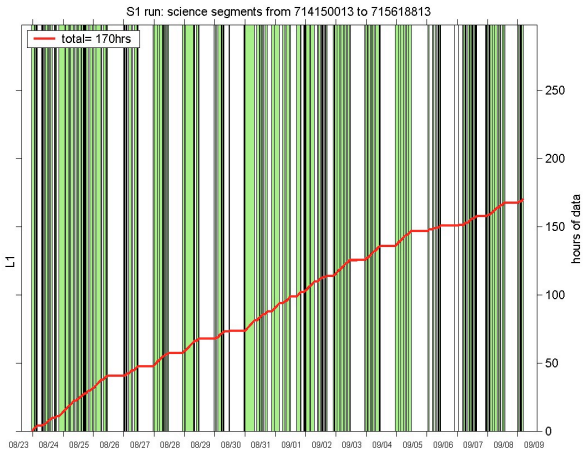
Aug – Sept 02



In-Lock Data from S1







Livingston 4km: 170 hrs of lock

Hanford 4km: 235 hrs of lock

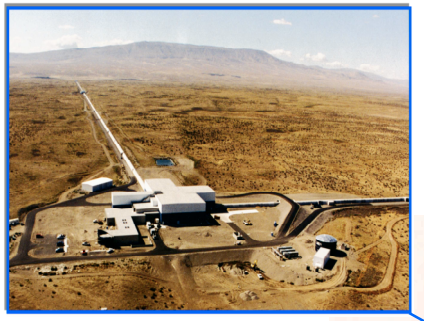
Hanford 2km: 298 hrs of lock

Triple Coincidence: 95.7 hrs →

LIGO Laboratory

Growing International Network of GW Interferometers

LIGO-LHO: 2km, 4km



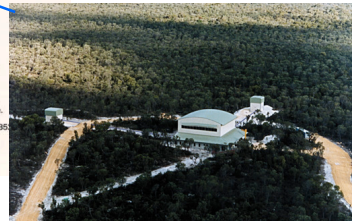
GEO: 0.6km

VIRGO: 3km

TAMA: 0.3km

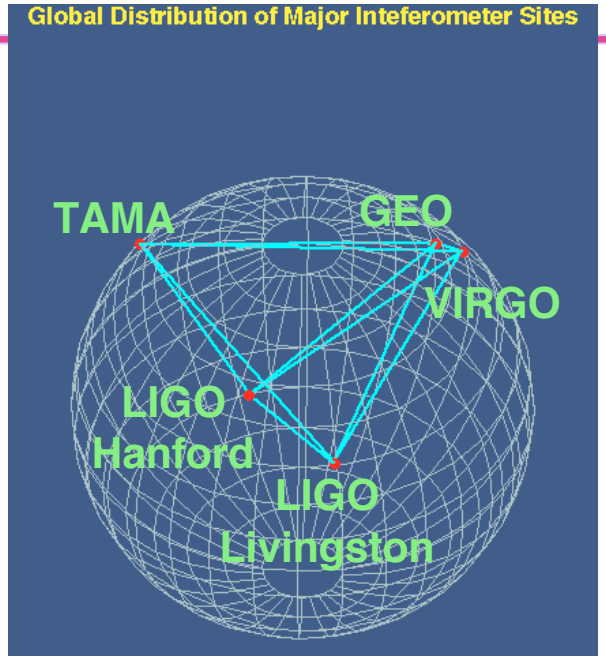
LIGO-LLO: 4km

AIGO: (?)km



LIGO Laboratory

Event Localization With An Array of GW Interferometers



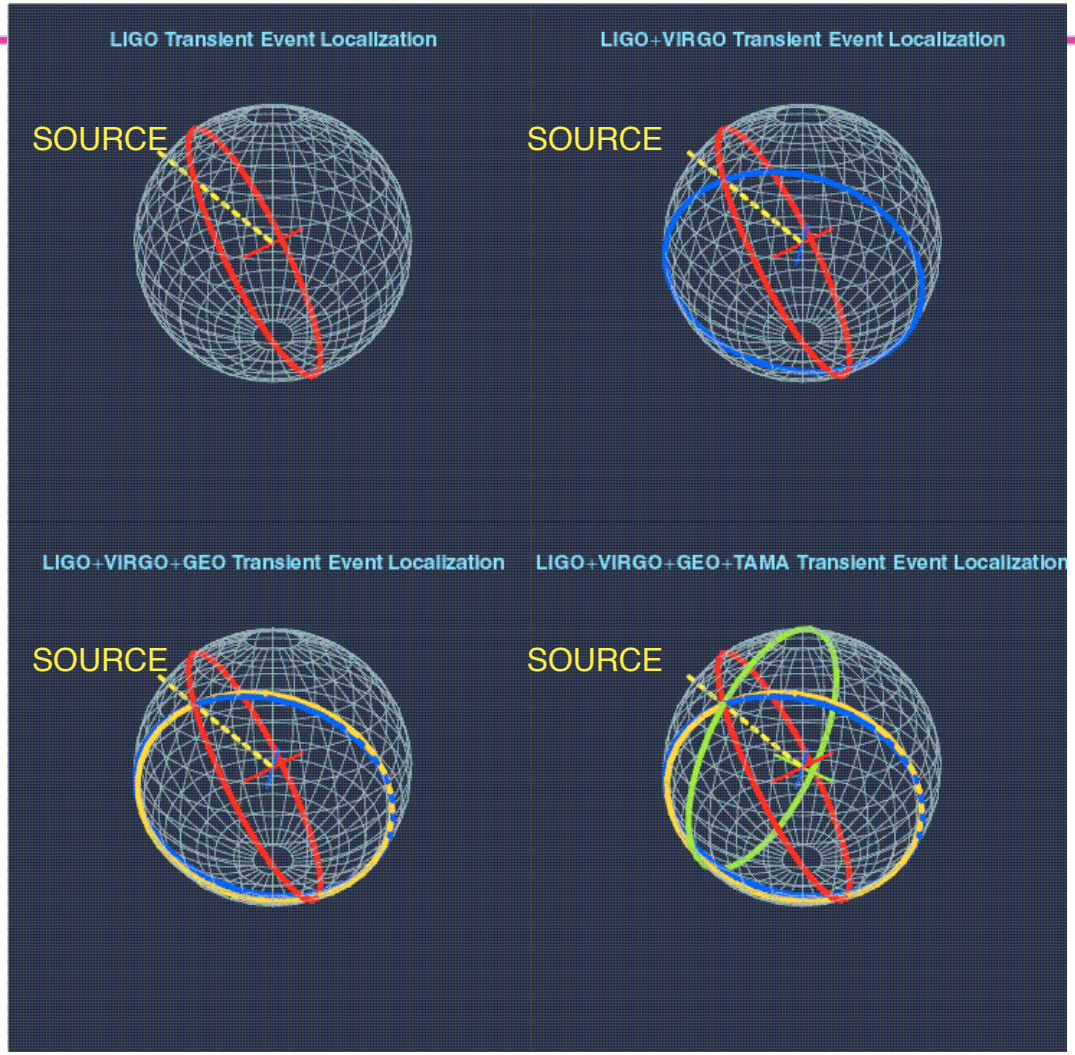
$\cos\theta = \delta t / (c\, D_{12})$

$\theta\theta \sim 0.5$ deg

$\delta L = \delta t\, c$

θ

1 2



LIGO Science Has Started

- LIGO had extremely successful first science run this summer!
  - » LSC Upper Limits Groups currently carrying out the data analysis!
- LIGO is taking its first steps to providing new scientific insight into the workings of the Universe.
- ***LIGO is taking and analyzing data***
  - » *Tier 2 centers are active participants -- **we are in operations***
  - » *Network analysis will further increase the off-line computational needs for the collaboration*

Remainder of this talk

- How did we get where we are today?
- Needs for the collaboration
  - » LIGO Laboratory
  - » Tier 2 Centers outside the Laboratory
- Issues and Questions

LIGO and the LIGO Scientific Collaboration

- 1989 - LIGO originally proposed. Scope provides for capability to look at the data at the proposing institutions
  - » Caltech
  - » MIT
- 1996 - NSF convenes Special Emphasis Panel on the Long Range Use of LIGO (Boyce McDaniel, chair)
  - » Recommends that a collaboration be formed around LIGO to enable the greatest benefit to be derived from this national resource
- 1997 - The LIGO Scientific Collaboration (LSC) formed in response to McDaniel Report
  - – Open collaboration (differs from the HEP model)
  - – Today: 35+ institutions world-wide, 400+ collaborators
  - – ***Implicit increase in LIGO Laboratory's role in data analysis support, data distribution & archiving***

LIGO and the LIGO Scientific Collaboration

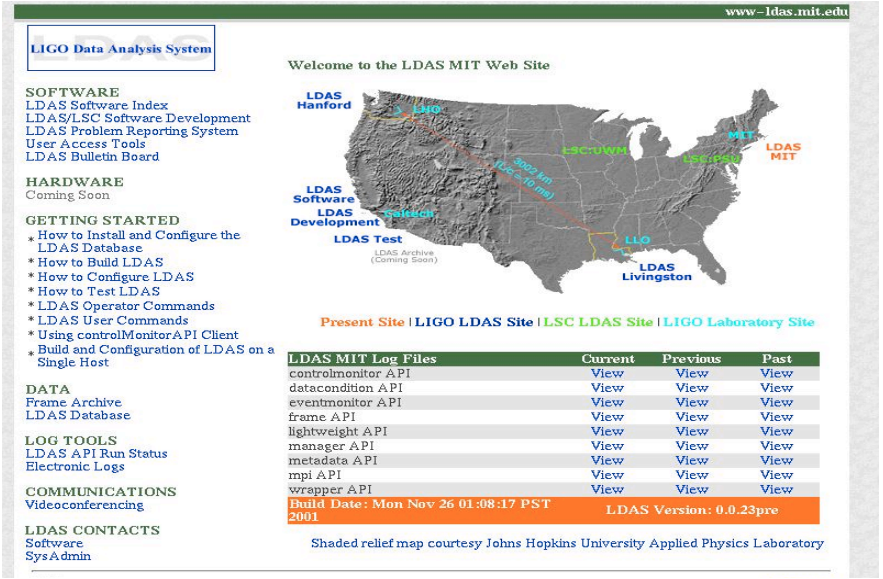
- 2000 - UWM receives MRI Grant -
  "Development of a high-capacity data analysis system for LIGO gravitational wave detection"
  - » Anticipates grid initiatives within NSF/MPS
  - » Beginning of "grid era" for LIGO and LSC
  - » Limited access resource -- 0 FTEs for support
- 2000 - GriPhyN (ITR2000)
  - » C/S + Applications R&D, prototyping
  - » 2 FTEs for C/S applications development (1 ea. at UWM, CIT)
  - » 1 FTE for outreach at UTB
- 2001 - iVDGL (ITR2001)
  - » 2 ea. Tier 2 centers for LSC
    - – Enhance UWM MRI facility, greenfield buildup at PSU
    - – Port, install GriPhyN deliverables to Tier 2 centers
  - » 0.6 FTEs for site support at UWM, PSU
  - » 2 FTEs - postdoc+GRAs for applications integration
  - » 0 FTEs at LIGO Laboratory

LIGO and the LIGO Scientific Collaboration Needs

- LIGO is now in data production
  - » Scientifically interesting data are flowing
  - » Tension between grid developments and scientific output
  - » Laboratory is trying to balance both demands
    - – We estimate ***more than 1 and as much as 2 FTEs (Caltech only)*** are devoted to supporting collaboration-wide activities in grid computing
      - Not presently in scope of Laboratory personnel
      - Effort provided “on the side”, “nights and weekends”
        - » Cannot continue indefinitely
- Continuing effort is needed to increase or maintain network connectivity for Caltech, MIT, LHO, and LLO to support grid computing and high-speed data replication to Tier-2 centers
- The proliferation of independent computer centers is a significant ***brain-drain*** from scientific progress in the field
  - » Scientists are being used as systems administrators, programmers

LIGO Laboratory Facilities @ MIT

- Tier-2 Laboratory and LSC facility
- Operational as of summer 2001
  - » 32 nodes LDAS cluster
- H/W and S/W support: Caltech
  - » start up and operations
- Current activities (0.5 FTE)
  - » Support LIGO eng/sci analyses of Upper Limits Groups (~15 users)
  - » H/W , S/W troubleshooting and commissioning (local)
- Expected growth: x2/x3
  - » Match data volume and computational needs of currently supported analyses
  - » Become seriously involved in development



LIGO and the LIGO Scientific Collaboration Needs

- LIGO is not unique in this regard
- At the ADASS (Astronomical Data Analysis Software and Systems) conference this past October:
  - » Kent Blackburn (LIGO SW System Architect) attended a talk covering the use of the grid in astronomy. Speaker asserted: "adding Grid Technology to your site would require ***an additional system administrator***."
    - – Experience confirmed by other groups in the discussion which followed the talk.
    - – There is a lot of administration associated with a working Grid installation which was at this early stage in the Grid Paradigm development compounded by the lack of maturity in the associated tools

ITR 2003

Proposal to NSF for iVDGL Operations

- LSC Computing Committee organizing a collaboration-wide effort to request funding to operate LSC Tier 2 centers for LIGO Science Run(s)
  - » *ITR2000: GriPhyN ->C/S + Applications R&D, prototyping*
  - » *ITR2001: iVDGL-> Center buildup (2 for LSC), port, install GriPhyN deliverables to Tier 2 centers*
  - » ITR2003: manpower to operate centers ***to do the science***
    - – Systems administration
    - – Help desk, 7x24 operations support
- Is the LIGO and LSC need well-matched to the ITR solicitation?

The UWM LIGO Data Analysis Center

Alan Wiseman

University of Wisconsin -- Milwaukee

UWM LSC Computing Facilities

Medusa cluster - designed for GW analysis work

- 296 nodes, each with
  - » 1 GHz Pentium III
  - » 512 MBytes memory
  - » 100baseT Ethernet
  - » 80 Gbyte disk
  - » on-board "health" monitoring
- UPS power
- Fully-meshed switch

Storage: 22 TBytes

CPU: 296 Gflops

Completed in August 2001

NSF MRI and UWM matching funds

Overview of UWM Tier 2 Center

- Goals:
  - » Full service, 24-7 LIGO Data analysis Center
  - » Data distribution center for LSC
  - » Full participant in deployment of "grid tools"
- Status:
  - » Operational: Currently being used for Upper Limit analysis
    - – LDAS is installed.
    - – Data Monitoring Tool is being installed
  - » Cluster is used "around-the-clock"
  - » Over 30 "outside" LIGO user accounts
  - » Non-LSC, Grid users: 50,000 jobs (CPU decades). More planned.
  - » Houses approximately 20TBytes of LIGO Data
  - » LIGO Data Replicator (LDR) Grid tool is being used
    - – Moves 2-20 Mega Bytes/sec between CIT/UWM/(Other Sites)
    - – Exceeds real-time data taking rate

History and Experience

- Cluster built with MRI funds: $450K NSF and $150K UWM Match
  - » Designed as quick-turnaround, prototyping facility
  - » Modest plan for support of external users
  - » Not designed as a 24-7 bullet-proof system
  - » Designed before grid computing had entered LSC model.
- Then paradigm shift to distributed computing: grid, Tier-N centers
  - » Cluster became a major computing engine in a Tier-2 center
    - – 24/7 system
    - – More (more impatient) external users: Upper Limit deadlines
    - – Broader range of software and hardware tool
      - LDAS Front end hardware
      - Data Monitoring Tool
      - Grid tools (condor, port-forwarding, flocking,
    - – Data storage and distribution
      - Faculty, postdocs in close collaboration with iVDGL facilities work group. Doing Development work.

History and Experience

- Change in role Required:
  - » Hardware upgrade:
    - – gone smoothly (at least affordably)
    - – UPS upgrade
    - – Switch reliability
    - – LDAS server upgrade (In progress)
    - – Major Storage upgrade (in progress)
      - Sufficient money to make intelligent choices
  - » Personnel upgrade for continuous operation:
    - – can't afford it.

Reality

- Operation of a center of this scale requires full time system administration support
  - » We have great (under)graduate students, but …
    they cannot (and should not) manage a 24-7 computing system
  - » Faculty, postdocs excellent for development efforts
    - – Not viable for maintaining 24-7 operations.
    - – Faculty (me) spending 30-50% doing sys-admin work. Brain drain.
  - » Currently employ one full-time sys-admin [Using a combo of yr1&2 funds]
  - » One is not enough!! Two per center.

Reality

- Current iVDGL funding is insufficient to sustain full time sysadmin support:
  - » iVDGL Sysadmin (technical) support over the next five years:

Year:	1	2	3	4	5
Sys-admin Mo:	4	5	4	1	0 (20%)

The Penn State LIGO Data Analysis Center

Lee Samuel Finn

Penn State

LIGO-G020511-02-E

Overview

- "Greenfield" Tier 2 Center
  - » Served by LIGO Tier 1 Center
  - » On-par with LHO, LLO, MIT, UWM
  - » Serves: LIGO
- Goal: "Full service" Analysis & Development LIGO Data Analysis Center
  - » Applications
    - – Analysis, detector characterization, Monte Carlo, simulations
  - » Software systems support
    - – LDAS, DMT, other tools (e.g., matlab, standalone C, C++, analysis tools)
- Configuration: "Clone" LIGO Lab facility
  - » Minimize resources, including sweat equity, spent adapting analysis tools to local customizations
  - » Maximize inter-site operability, resources available for supporting LIGO analysis activities
- Status
  - » Small (12 node) pathfinder system purchased to evaluate networking options & gain experience with LDAS configuration, operations
  - » Approaching decision point on h/w for DMT support

Personnel Resources

- iVDGL-funded (4 yrs)
  - 1 Postdoc: hired (1 Nov start)
  - 50% Sysadmin: center support (searching)
- PSU-contributed to establish Center
  - (Partnership with HPC group)
  - 0.1 FTE Director, HPC group
  - 0.4 FTE Sr Rsrch Prgmr
  - 0.4 FTE Rsrch Prgmr
  - 1 FTE (CS/EE) graduate student (globus/iVDGL focus)
  - 0.2 FTE web administrative support

Reality: iVDGL funding insufficient for Center support

- Large scale computing for production analysis requires dedicated, professional support
- What is required to support center mission?
  - » 1 FTE h/w, o/s systems administration
    - – 70 nodes, 30 TB storage (6 mo RDS)
  - » 1 FTE s/w applications system administration
    - – Maintain & support ldas, dmt, database, other s/w systems configurations & upgrades
    - – Liaison with other tier 2, tier 1 centers (data exchange, database federation, etc.)
  - » 0.5-1 FTE at tier 1 center
    - – User support/help desk/liaison/development

Summary

- "Greenfield" Tier 2 Center
  - » Goal: "Full service" Analysis & Development LIGO Data Analysis Center
- Configuration: "Clone" LIGO Lab facility
  - » Minimize resources, including sweat equity, spent adapting analysis tools to local customizations
  - » Maximize inter-site operability, resources available for supporting LIGO analysis activities
- Status
  - » Small (12 node) pathfinder system purchased to evaluate networking options & gain experience with LDAS configuration, operations
  - » Approaching decision point on h/w for DMT support
- Reality: iVDGL funding insufficient for Center support
  - » Required: 2 FTE IT professional at PSU, 0.5-1 FTE at tier 1 (LIGO/CIT)