NARIT

Big Data & HPC:

Astronomical Numerical Simulations

Alfredo J. Santillán

UNAM

21st century can be the era of the Big Data.

"In 2017, the Economist claimed that data replaced oil as the world's most valuable source."

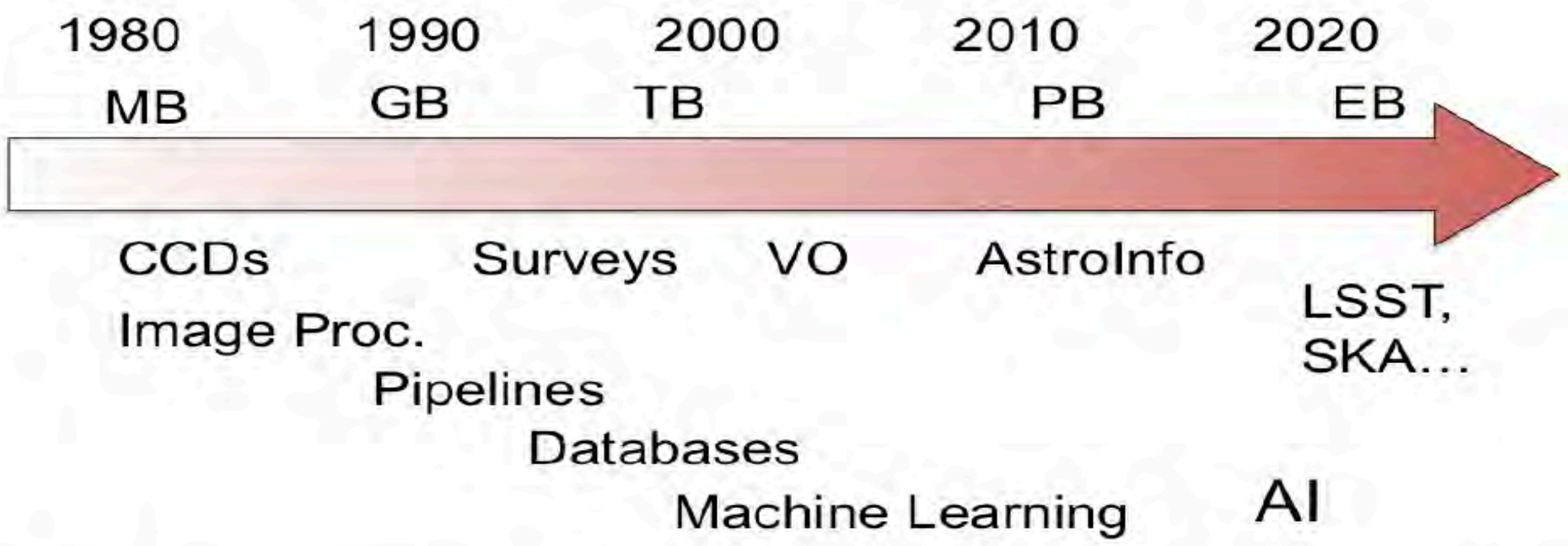
definition of the characteristics of Big Data “10 V”:
Volume, Variety, Velocity, Veracity, Validity, Value, Variability, Venue, Vocabulary & Vagueness

Some interesting statistics about Big Data.

- By 2020, there will be around 40 ZB.
- 90% of all data has been created in the last 2 years.
- Today a person may need approximately 181 Myr to download all the data from the internet (44Mbps/0.55 ZB).
- In 2017, the Economist claimed that data replaced oil as the world's most valuable source.
- By 2020, every person will generate 1.7 MB in just a second.
- Internet users generate about 2.5 quintillion bytes of data each day.
- Automated analytics will be vital to Big Data by 2020.

The Evolving Data-Rich Astronomy

An example of a "Big Data" science driven by the advances in computing/information technology



Key challenges: data heterogeneity and complexity

Djorgovski

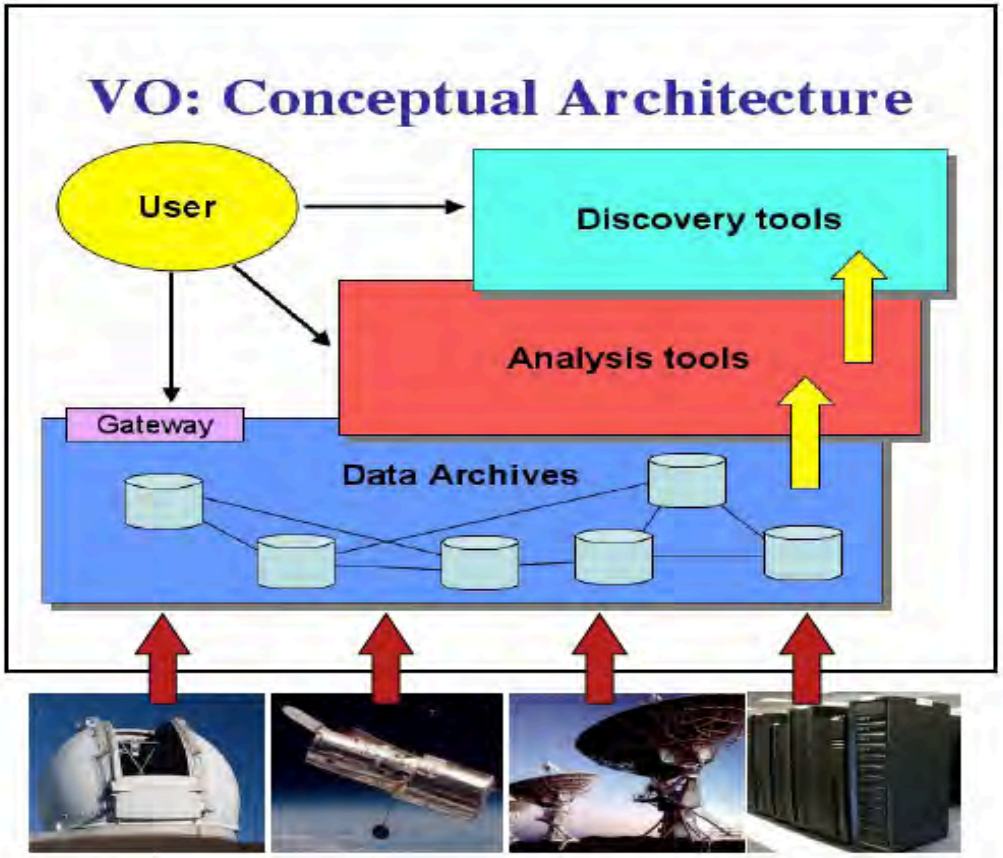
The Virtual Observatory Concept

- A complete, dynamical, distributed, open ***research environment for the new astronomy with massive and complex data sets***

– Provide and federate content (data, metadata) services, standards, and analysis/compute services

– Develop and provide data exploration and discovery tools

– A successful example of an e-Science /Cyber-Infrastructure



Djorgovski

DEVELOPMENT THROUGH RADIO ASTRONOMY
GLOBAL NETWORK

NARIT

Astro
Grid

ArVO
Armenian Virtual Observatory

ASVO
ALL SKY VIRTUAL OBSERVATORY

BRAVO

China-VO

ChiVO
CHILEAN
VIRTUAL
OBSERVATORY

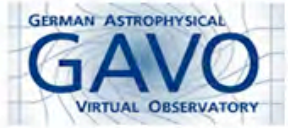
CVO

esa
European Space Agency

France

IVOA

EURO VO


GERMAN ASTROPHYSICAL
GAVO
VIRTUAL OBSERVATORY

HVO
Hungarian Virtual Observatory

JVO

NOVA

RVO

NRF
National Research Foundation
SA3
South African Astroinformatics Alliance

SVO

UKR
VO

USVOA

VObs.it

vo-i

Database Generators

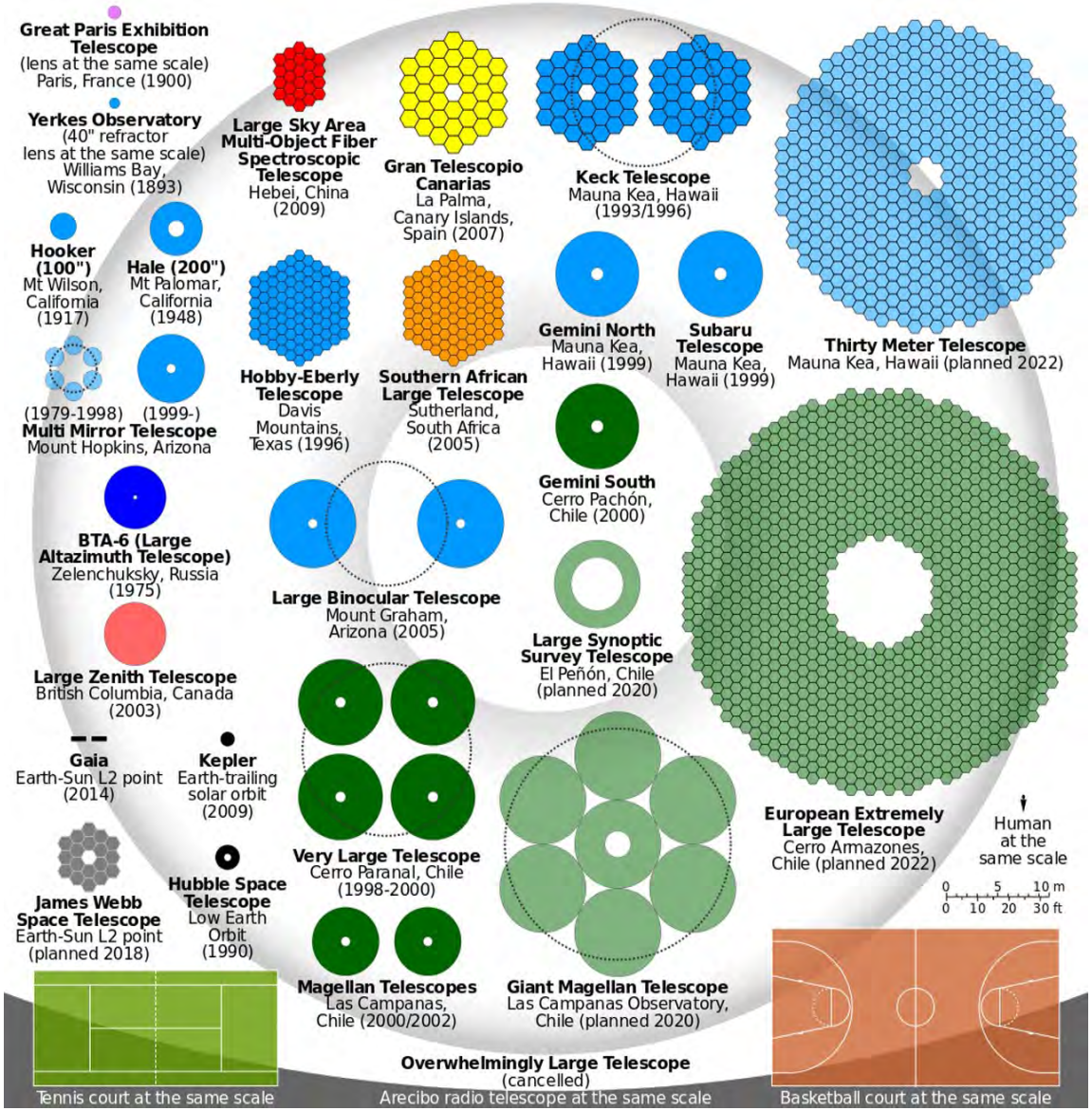
- Telescopes
  - Ground
    - SKA, ALMA, SDSS, LSST…
  - Space
    - HST, Chandra, Spitzer…
- Numerical Simulations
  - Millennium Simulation project
  - IllustrisTNG project
  - …

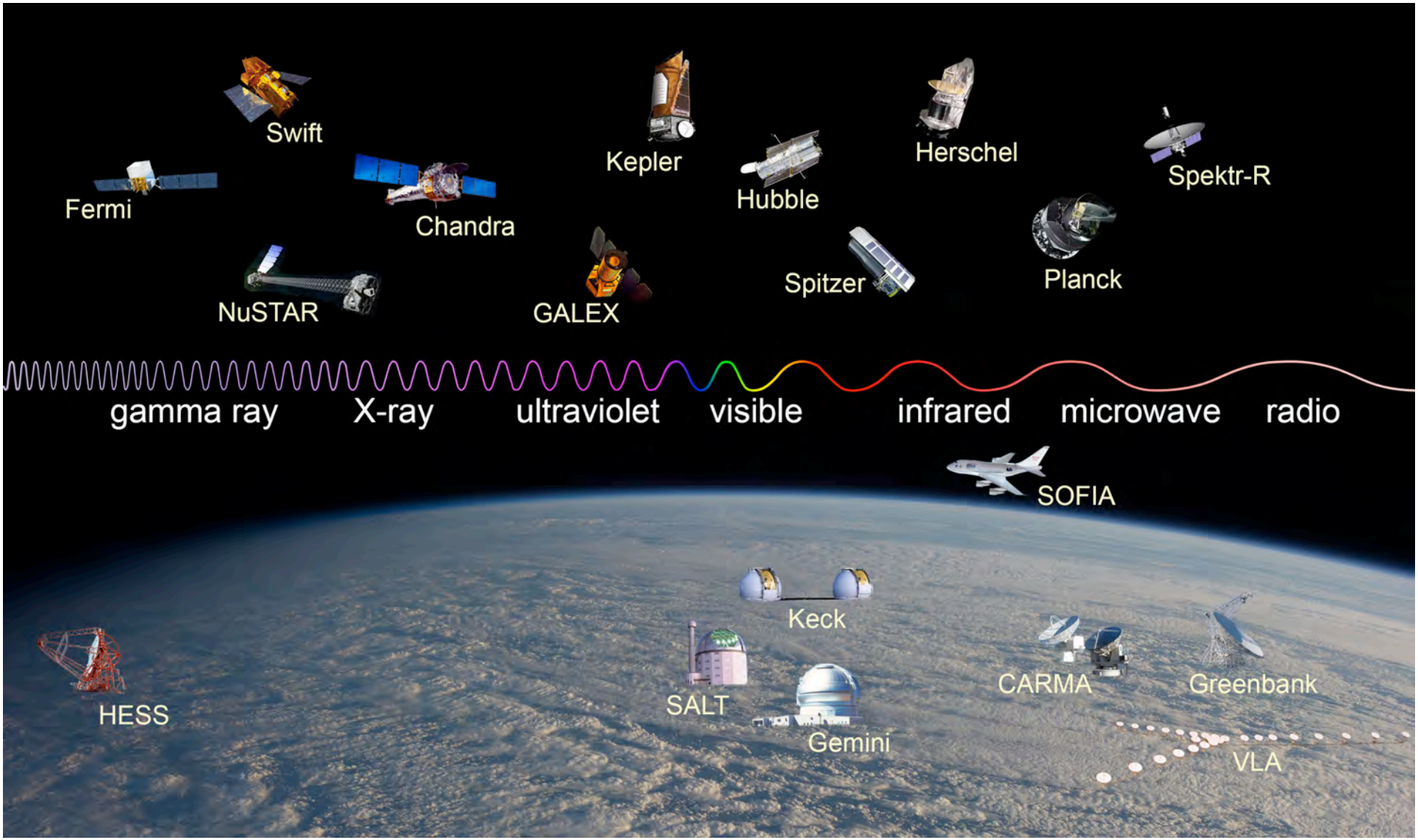
ASTRO

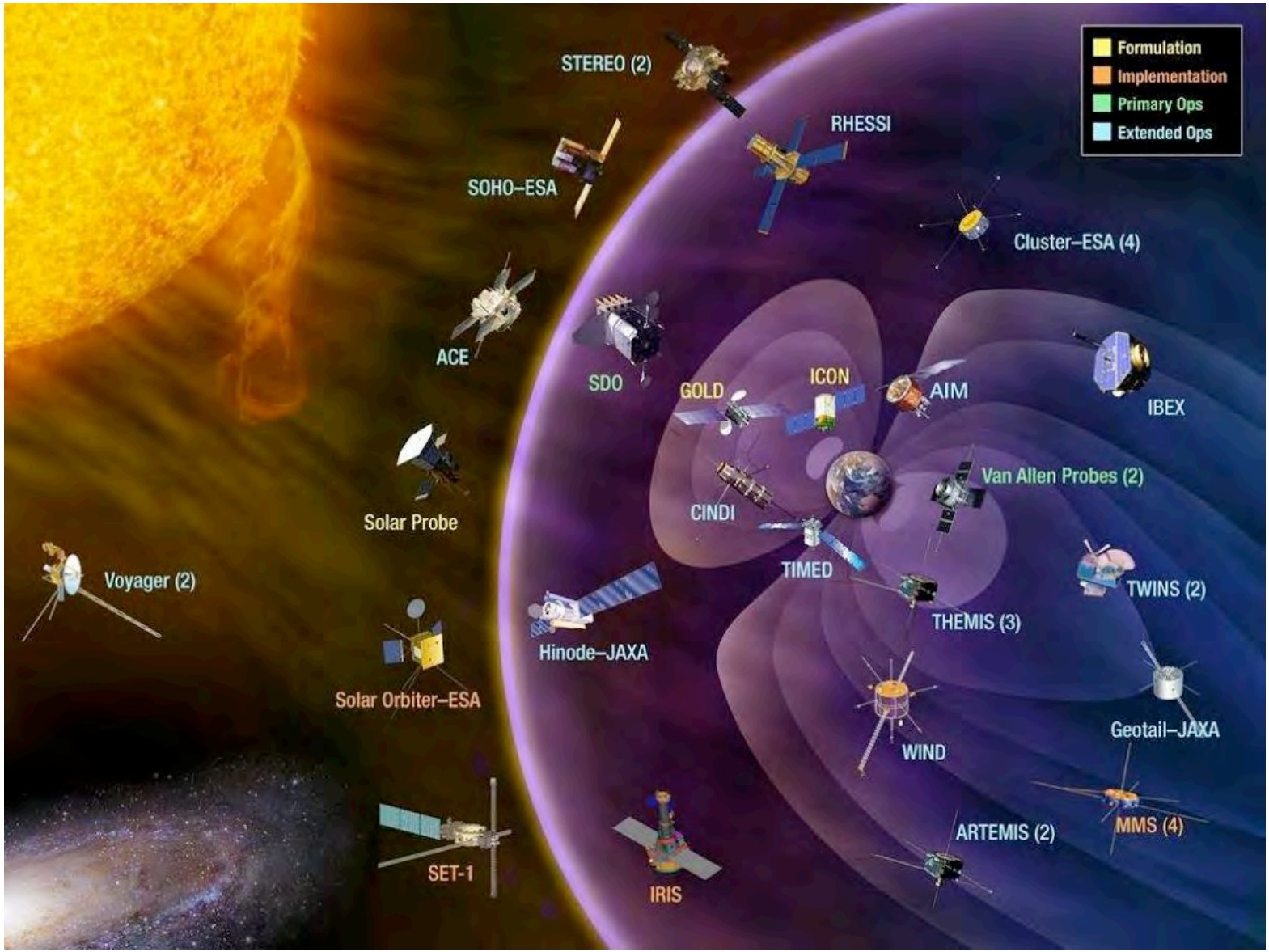
- Biology
- Chemistry
- Statistics
- Informatics
- …

Astrostatistics is an interdisciplinary field of astronomy and statistics that applies statistics to the study and analysis of astronomical data.

Astroinformatics is an interdisciplinary field of astronomy and informatics that uses information/communications technologies to solve the big data problems faced in astronomy.


Great Paris Exhibition Telescope
(lens at the same scale)
Paris, France (1900)
Yerkes Observatory
(40" refractor
lens at the same scale)
Williams Bay,
Wisconsin (1893)
Hooker (100")
Mt Wilson,
California
(1917)
Hale (200")
Mt Palomar,
California
(1948)
(1979-1998)
(1999-)
Multi Mirror Telescope
Mount Hopkins, Arizona
BTA-6 (Large Altazimuth Telescope)
Zelenchuksky, Russia
(1975)
Large Zenith Telescope
British Columbia, Canada
(2003)
Gaia
Earth-Sun L2 point
(2014)
Kepler
Earth-trailing
solar orbit
(2009)
James Webb Space Telescope
Earth-Sun L2 point
(planned 2018)
Hubble Space Telescope
Low Earth
Orbit
(1990)
Large Sky Area Multi-Object Fiber Spectroscopic Telescope
Hebei, China
(2009)
Gran Telescopio Canarias
La Palma,
Canary Islands,
Spain (2007)
Keck Telescope
Mauna Kea, Hawaii
(1993/1996)
Hobby-Eberly Telescope
Davis
Mountains,
Texas (1996)
Southern African Large Telescope
Sutherland,
South Africa
(2005)
Gemini North
Mauna Kea,
Hawaii (1999)
Subaru Telescope
Mauna Kea,
Hawaii (1999)
Thirty Meter Telescope
Mauna Kea, Hawaii (planned 2022)
Gemini South
Cerro Pachón,
Chile (2000)
Large Binocular Telescope
Mount Graham,
Arizona (2005)
Large Synoptic Survey Telescope
El Peñón, Chile
(planned 2020)
Very Large Telescope
Cerro Paranal, Chile
(1998-2000)
Magellan Telescopes
Las Campanas,
Chile (2000/2002)
Giant Magellan Telescope
Las Campanas Observatory,
Chile (planned 2020)
European Extremely Large Telescope
Cerro Armazones,
Chile (planned 2022)
Human at the same scale
0 5 10 m
0 10 20 30 ft
Overwhelmingly Large Telescope
(cancelled)
Tennis court at the same scale
Arecibo radio telescope at the same scale
Basketball court at the same scale


Swift
Kepler
Herschel
Spektr-R
Fermi
Chandra
Hubble
Planck
Spitzer
NuSTAR
GALEX
gamma ray
X-ray
ultraviolet
visible
infrared
microwave
radio
SOFIA
Keck
HESS
SALT
Gemini
CARMA
Greenbank
VLA


Formulation
Implementation
Primary Ops
Extended Ops
STEREO (2)
RHESSI
SOHO–ESA
Cluster–ESA (4)
ACE
SDO
GOLD
ICON
AIM
IBEX
Van Allen Probes (2)
CINDI
Solar Probe
TIMED
TWINS (2)
Voyager (2)
THEMIS (3)
Hinode–JAXA
Solar Orbiter–ESA
Geotail–JAXA
WIND
MMS (4)
ARTEMIS (2)
SET-1
IRIS

Actualmente el proyecto ASKAP (*Australian Square Kilometre Array Pathfinder*) produce 7.5 TB/s de una muestra de datos de imágenes, para 2025 esta tasa se incrementará a 750 TB/s ~ 25 ZB/año.

Catedral de Turku, Finlandia

Torre Redonda de Copenhague, Dinamarca

Cristo Redentor de Río de Janeiro, Brasil

A depiction of what the completed LSST observatory will look like atop El Peñon summit, Chile.

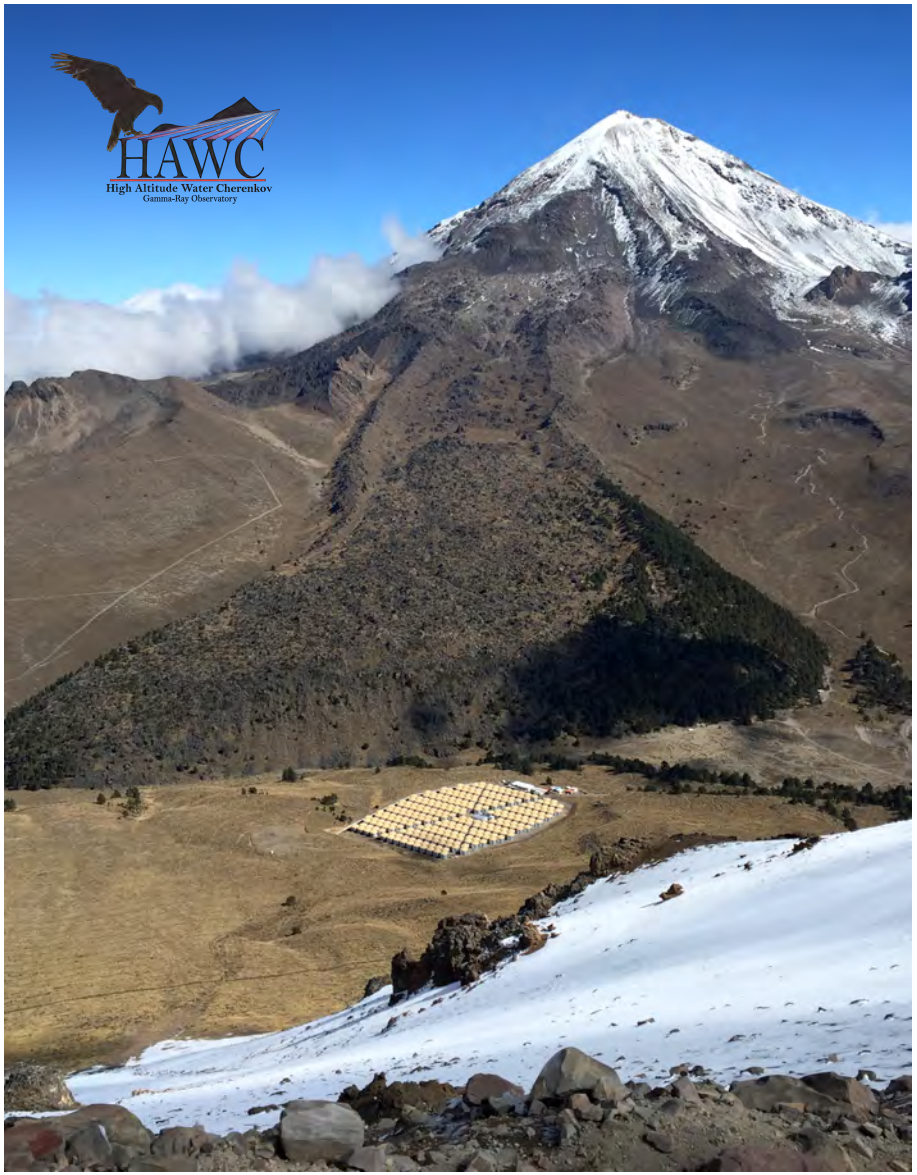
Large Synoptic Survey Telescope (LSST)

Large Synoptic Survey Telescope

- Will be the largest digital instrument ever built for ground-based astronomy and will take pictures fast enough to capture the entire southern sky every three nights.
- The Hubble Space Telescope would need 120 years to image an equivalent area of sky.
- Imaging at this rate will generate about 15 TB of raw data per night and 30 PB over its 10-year survey life. (A petabyte is approximately the amount of data in 200,000 movie-length DVDs.) Even after processing, that's still a 15 PB store.

Sky Survey Projects	Data Volume
DPOSS (The Palomar Digital Sky Survey)	3 TB
2MASS (The Two Micron All-Sky Survey)	10 TB
GBT (Green Bank Telescope)	20 PB
GALEX (The Galaxy Evolution Explorer)	30 TB
SDSS (The Sloan Digital Sky Survey)	40 TB
SkyMapper Southern Sky Survey	500 TB
PanSTARRS (The Panoramic Survey Telescope and Rapid Response System)	~ 40 PB expected
LSST (The Large Synoptic Survey Telescope)	~ 200 PB expected
SKA (The Square Kilometer Array)	~ 4.6 EB expected

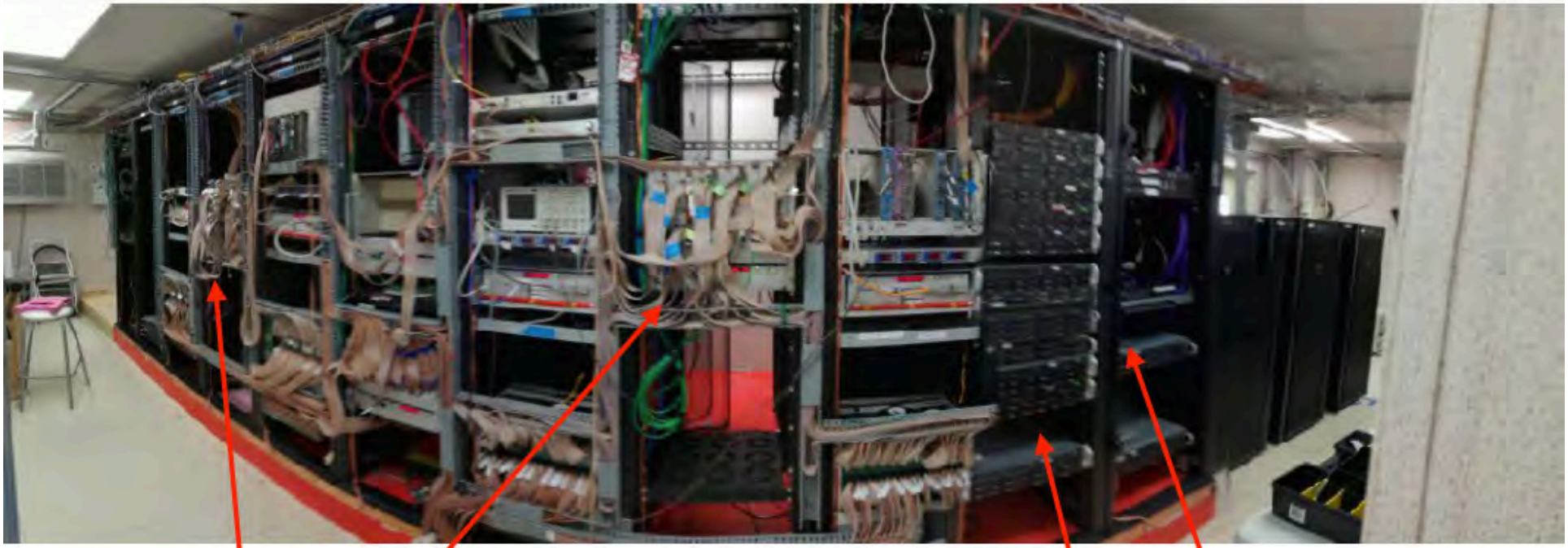
HAWC está ubicado dentro del Parque Nacional Pico de Orizaba, un parque nacional Mexicano. Dentro del parque se encuentra el Citlatepetl (o Pico de Orizaba), la montaña más alta de México con 5,610 m, y Sierra Negra, un volcán extinto con 4,600 metros ubicado a 7km al suroeste del Citlatepetl.

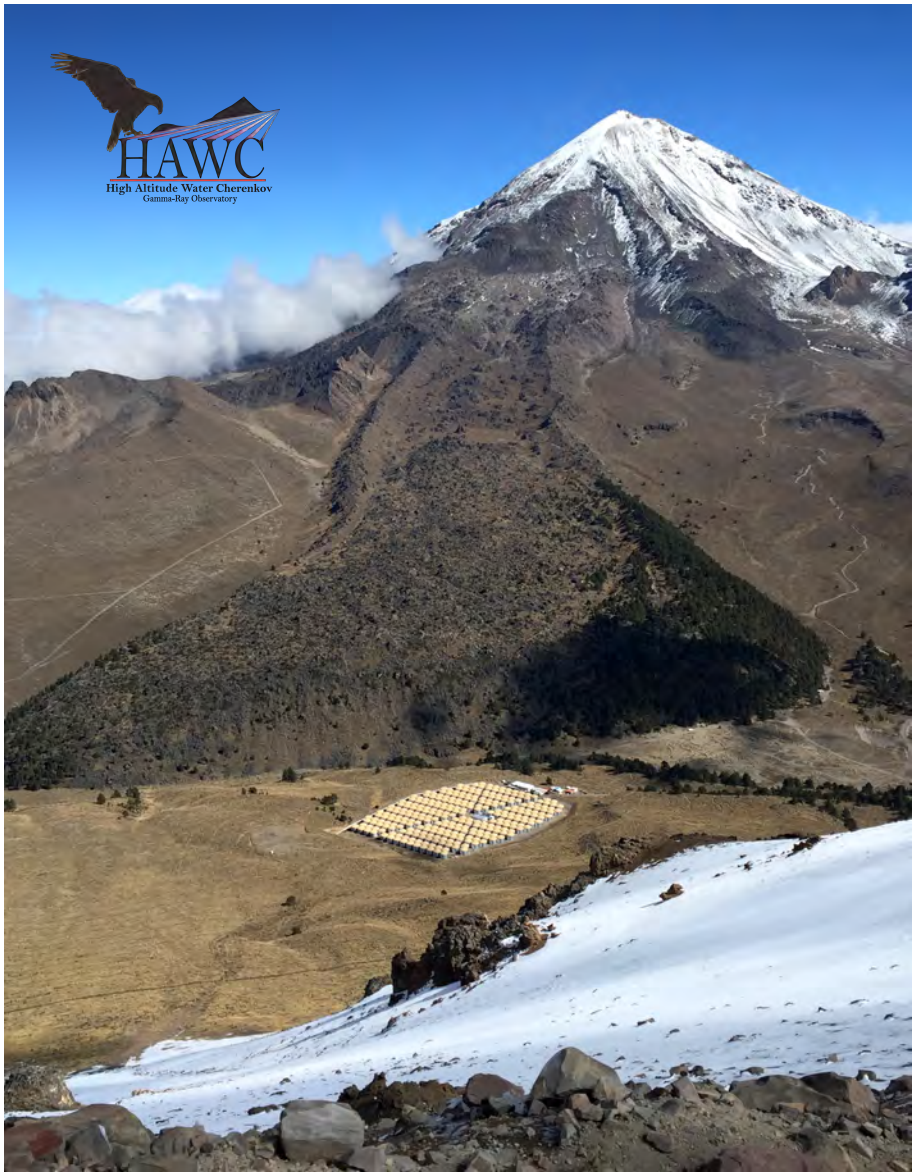


Data set.

- About 30TB of data in 10 years of operation.
- Upgrade will double data rate.
- 60TB in the next 10 years.
- Simulation set ~100TB.
- Reasonably stable.
- Software evolves.
- New simulations replace outdated ones.

The HAWC electronics room



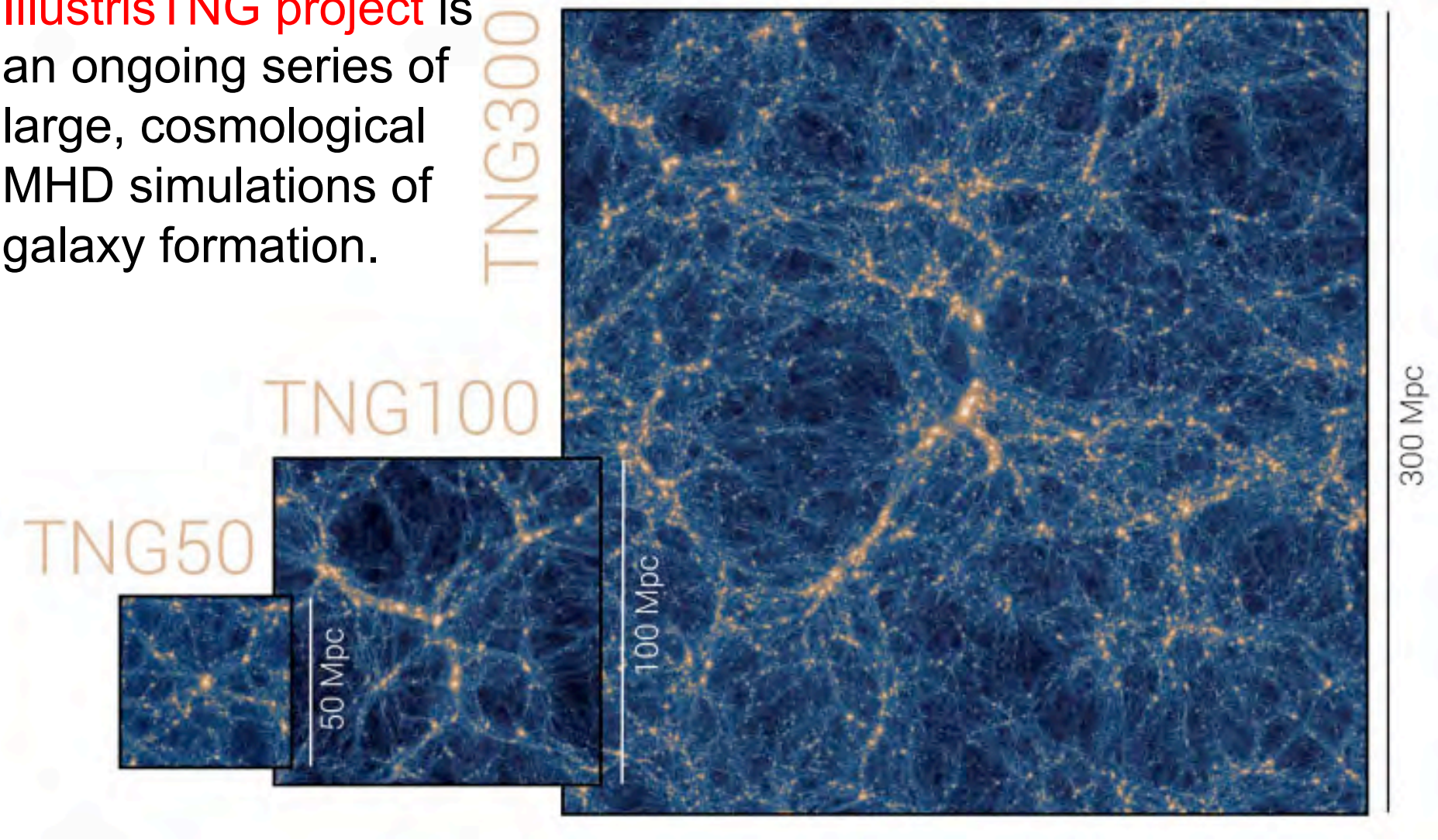


Storage.

- Data volume growing (data and simulation) ~700-1000 TB per year.
- We have already more 2 PB collected.
- Data sets of this size new for Mexico.
- Traditional super-computing CPU oriented.
- Possible to handle in an institute.
- Share experience and, if possible, equipment.
- Particle physics.
- Astronomy.
- Geosciences.
- Life sciences, bio-informatics.

- **Millennium Simulation:**
- Max Planck, Supercomputing Center, Garching, Germany.
- + than 10 billion particles to follow the dark matter distribution.
- Cubic region of the Universe over 60 Gpc on a side.
- 25 Tbytes of stored output.
- 20 million galaxies.
- They have implemented a Structured Query Language (SQL) server on these databases.
- VIRGO Consortium, UK, Germany, Japan, Canada &USA.

IllustrisTNG project is an ongoing series of large, cosmological MHD simulations of galaxy formation.



Simulation Name	$L_{\rm box}\,[Mpc]$	$N_{\rm DM}$	$m_{\rm DM}\,[M_\odot]$	$m_{\rm gas}\,[M_\odot]$	$N_{\rm snap}$	$N_{\rm Subfind}(z=0)$	Snaps	FoF	Subfind	SubLink	LHaloTree
TNG100-1	110.7	1820^3	7.5×10^6	1.4×10^6	100	4371211	✓	✓	✓	✓	✓
TNG100-1-Dark	110.7	1820^3	8.9×10^6	0	100	5012155	✓	✓	✓	✓	✓
TNG100-2	110.7	910^3	6.0×10^7	1.1×10^7	100	698336	✓	✓	✓	✓	✓
TNG100-2-Dark	110.7	910^3	7.1×10^7	0	100	757923	✓	✓	✓	✓	✓
TNG100-3	110.7	455^3	4.8×10^8	8.9×10^7	100	118820	✓	✓	✓	✓	✓
TNG100-3-Dark	110.7	455^3	5.7×10^8	0	100	116020	✓	✓	✓	✓	✓
TNG300-1	302.6	2500^3	5.9×10^7	1.1×10^7	100	14485709	✓	✓	✓	✓	✓
TNG300-1-Dark	302.6	2500^3	7.0×10^7	0	100	15724587	✓	✓	✓	✓	✓
TNG300-2	302.6	1250^3	4.7×10^8	8.8×10^7	100	2471022	✓	✓	✓	✓	✓
TNG300-2-Dark	302.6	1250^3	5.6×10^8	0	100	2411570	✓	✓	✓	✓	✓
TNG300-3	302.6	625^3	3.8×10^9	7.0×10^8	100	391144	✓	✓	✓	✓	✓
TNG300-3-Dark	302.6	625^3	4.5×10^9	0	100	372177	✓	✓	✓	✓	✓
Illustris-1	106.5	1820^3	6.3×10^6	1.3×10^6	134	4366546	✓	✓	✓	✓	✗
Illustris-1-Dark	106.5	1820^3	7.5×10^6	0	136	4872374	✓	✓	✓	✓	✓
Illustris-2	106.5	910^3	5.0×10^7	1.0×10^7	136	689785	✓	✓	✓	✓	✓
Illustris-2-Dark	106.5	910^3	6.0×10^7	0	136	735751	✓	✓	✓	✓	✗
Illustris-3	106.5	455^3	4.0×10^8	8.1×10^7	136	121209	✓	✓	✓	✓	✓
Illustris-3-Dark	106.5	455^3	4.8×10^8	0	136	111992	✓	✓	✓	✓	✗

Currently showing **18** runs with **2014** total snapshots and **651.1 TB** total data volume,
including **5,911,438,785** FoF groups, **5,772,665,381** Subfind groups, and **10,123,660,679,028** particles.

Big Data: Astronomical or Genomical?

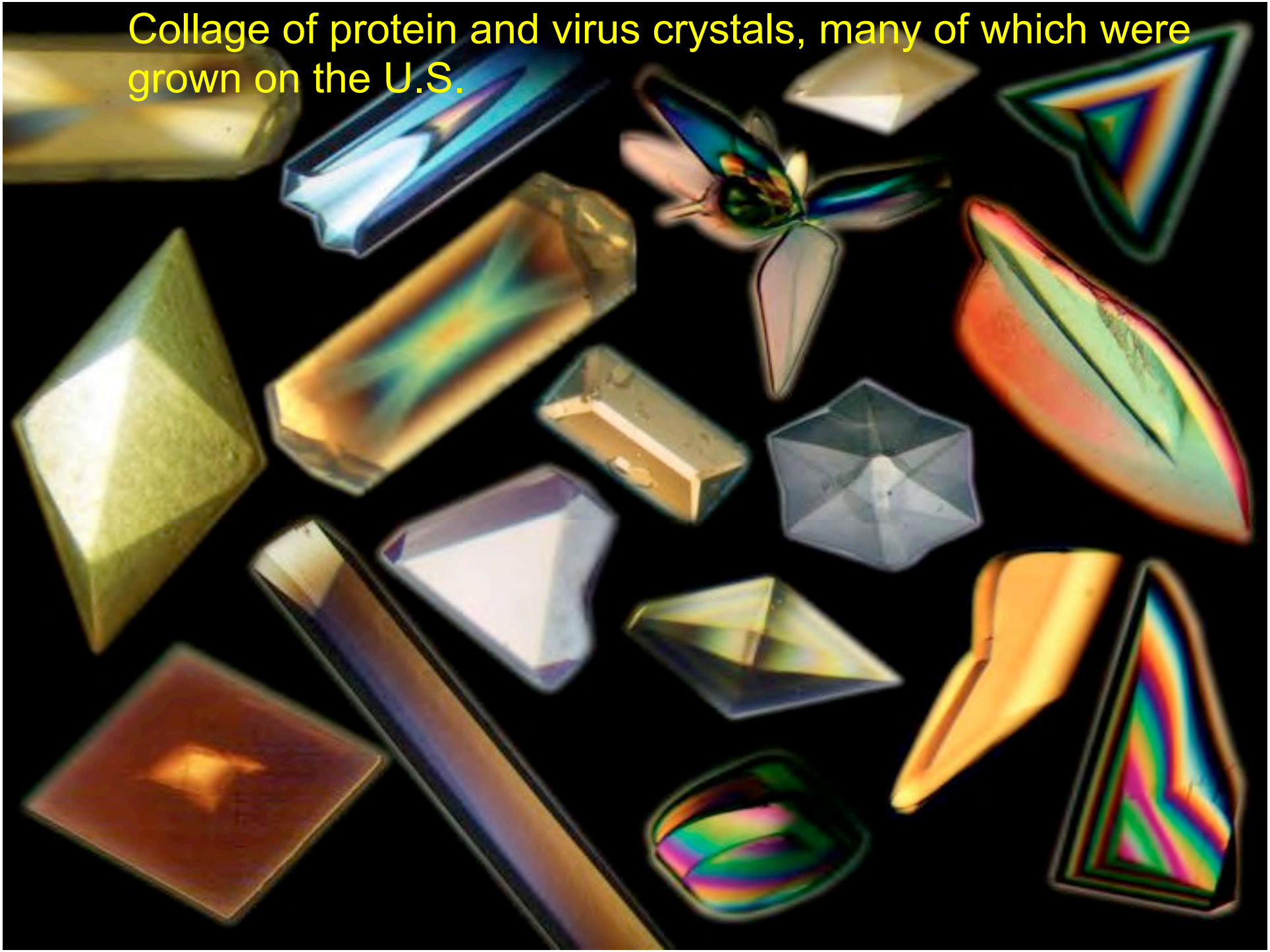
Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, Gene E. Robinson

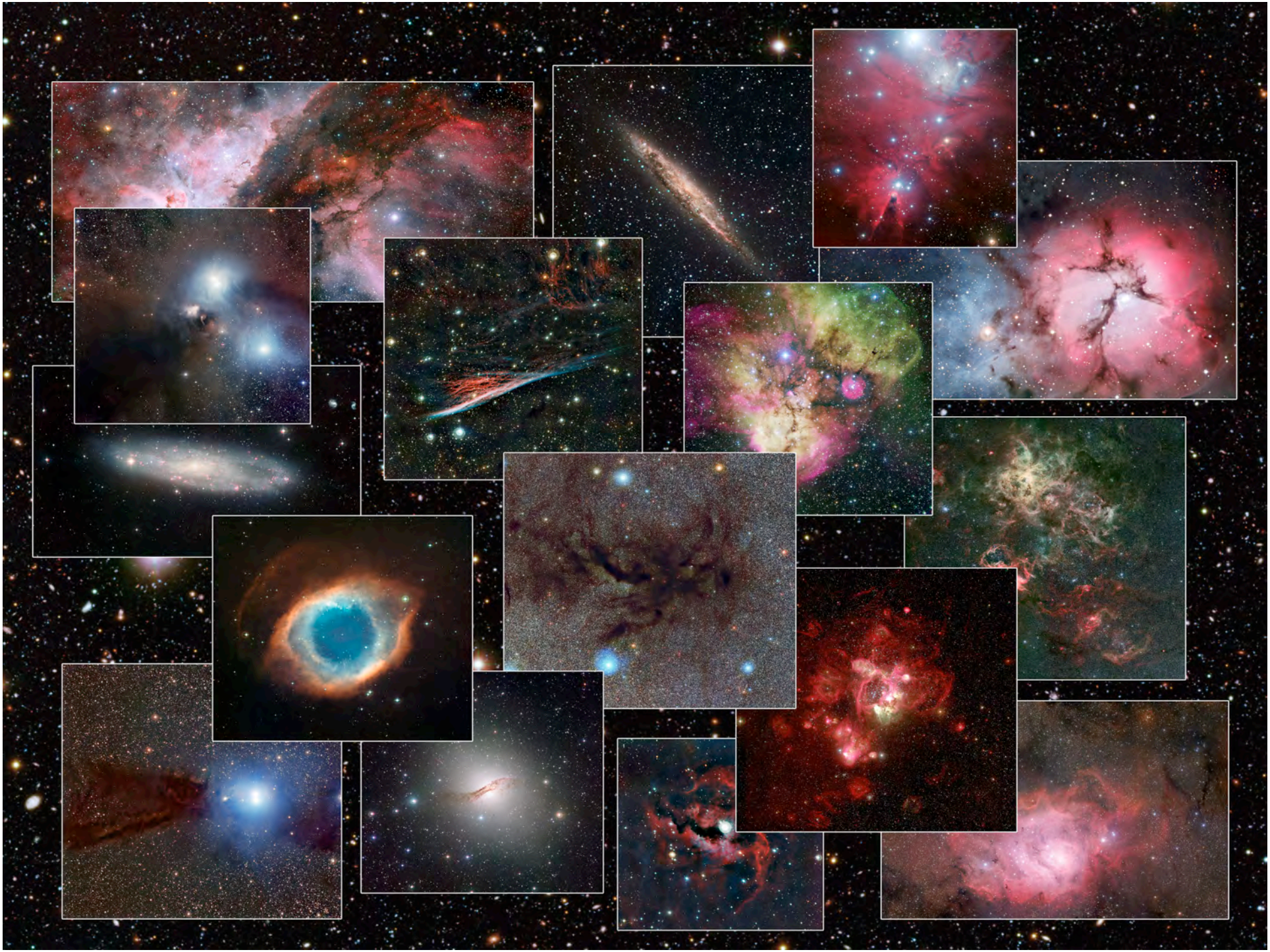
Published: July 7, 2015 • https://doi.org/10.1371/journal.pbio.1002195

Abstract
Data Acquisition
Data Storage
Data Distribution
Data Analysis
The Long Road Ahead
Supporting Information
Acknowledgments
References

Abstract

Genomics is a Big Data science and is going to get much bigger, very soon, but it is not known whether the needs of genomics will exceed other Big Data domains. Projecting to the year 2025, we compared genomics with three other major generators of Big Data: astronomy, YouTube, and Twitter. Our estimates show that genomics is a "four-headed beast"—it is either on par with or the most demanding of the domains analyzed here in terms of data acquisition, storage, distribution, and analysis. We discuss aspects of new technologies that will need to be developed to rise up and meet the computational challenges that genomics poses for the near future. Now is the time for concerted, community-wide planning for the "genomical" challenges of the next decade.


Collage of protein and virus crystals, many of which were grown on the U.S.



bacteria in petri dishes

Astronomy vs Genomics

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Micro vs Macro

Big Data

Brian May

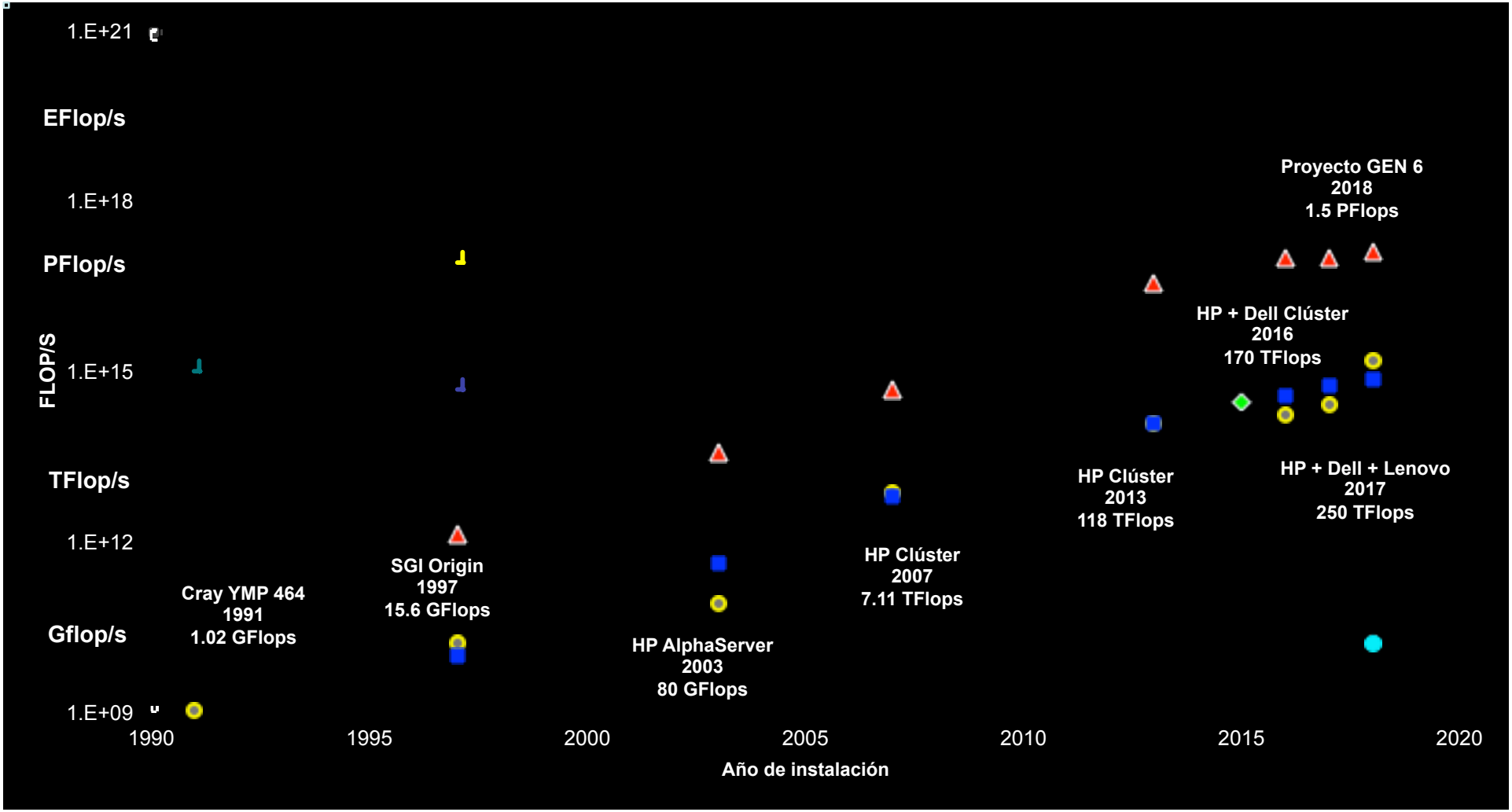
David Gilmour

Biotechnology & HPC

- Microbiology, Molecular biology, Biophysics.
- Genetics, Immunology, Neuroscience.
- Chemistry, Biochemistry, Physical chemistry.
- Bioprocess engineering, Screening, Bioinformatics.
- Microscopy and spectroscopy.
- Micro- and nanotechnology.
- Biomechanics.
- Bioelectronics.
- Aerosol research.
- Criminological applications.

Performance Development
10 EFlop/s
1 EFlop/s
100 PFlop/s
10 PFlop/s
1 PFlop/s
100 TFlop/s
10 TFlop/s
1 TFlop/s
100 GFlop/s
10 GFlop/s
1 GFlop/s
100 MFlop/s
Performance
1995
2000
2005
2010
2015
Lists
Sum
#1
#500

HPC UNAM


1.E+21
EFlop/s
1.E+18
PFlop/s
FLOP/S
1.E+15
TFlop/s
1.E+12
Gflop/s
1.E+09
1990
1995
2000
2005
2010
2015
2020
Año de instalación
Cray YMP 464
1991
1.02 GFlops
SGI Origin
1997
15.6 GFlops
HP AlphaServer
2003
80 GFlops
HP Clúster
2007
7.11 TFlops
HP Clúster
2013
118 TFlops
HP + Dell Clúster
2016
170 TFlops
Proyecto GEN 6
2018
1.5 PFlops
HP + Dell + Lenovo
2017
250 TFlops

Sin servicio 11:12 61%

Linpack
Benchmark

iPhone10,4-D201AP / 6 cores

Problem size: 500

Number of runs: 10

Multithread mode:

Run benchmark

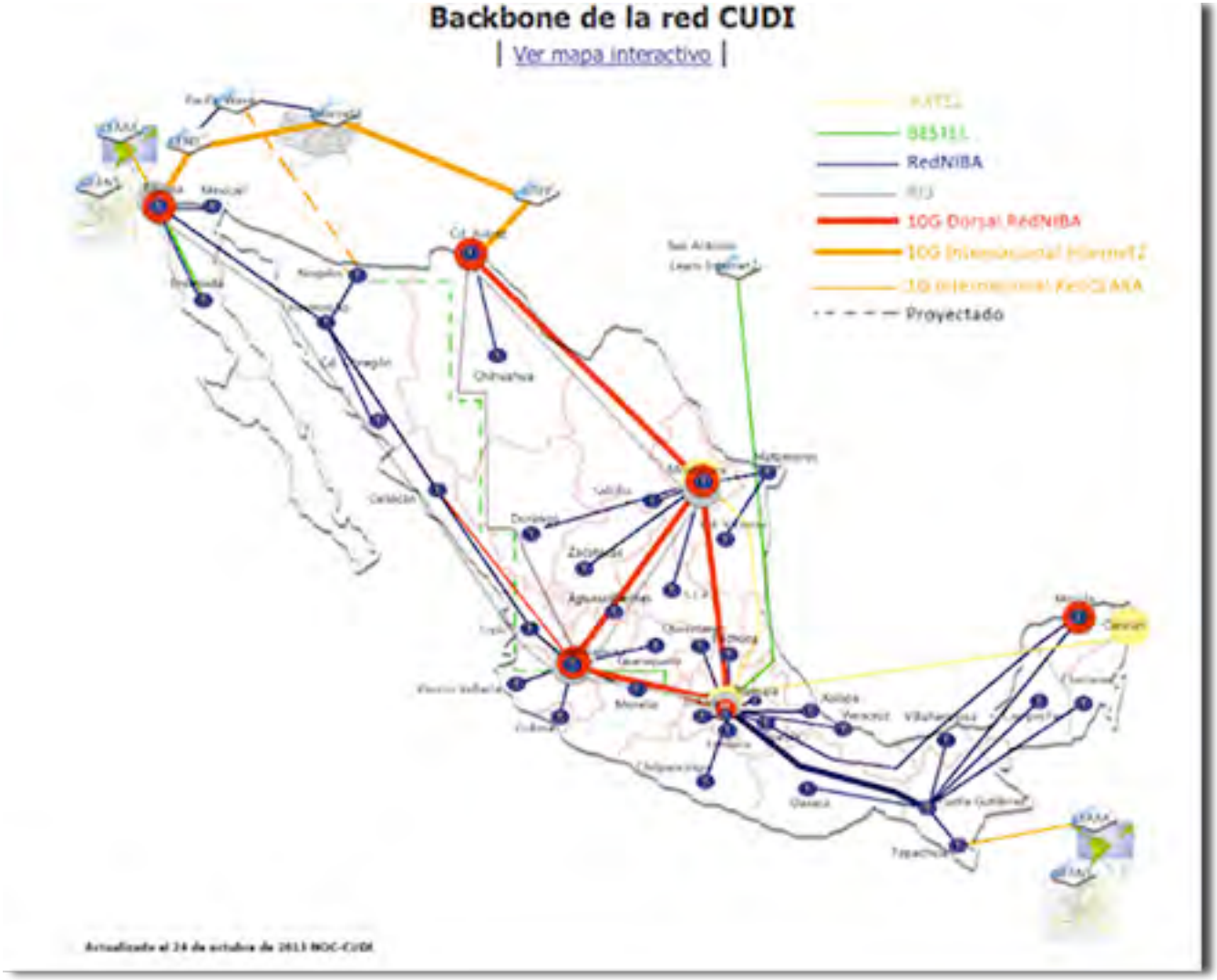
Run: #10
Mflop/s: 13743.12
Time: 0.0493
Norm Res: 5.1700
Precision: 2.22044605e-16

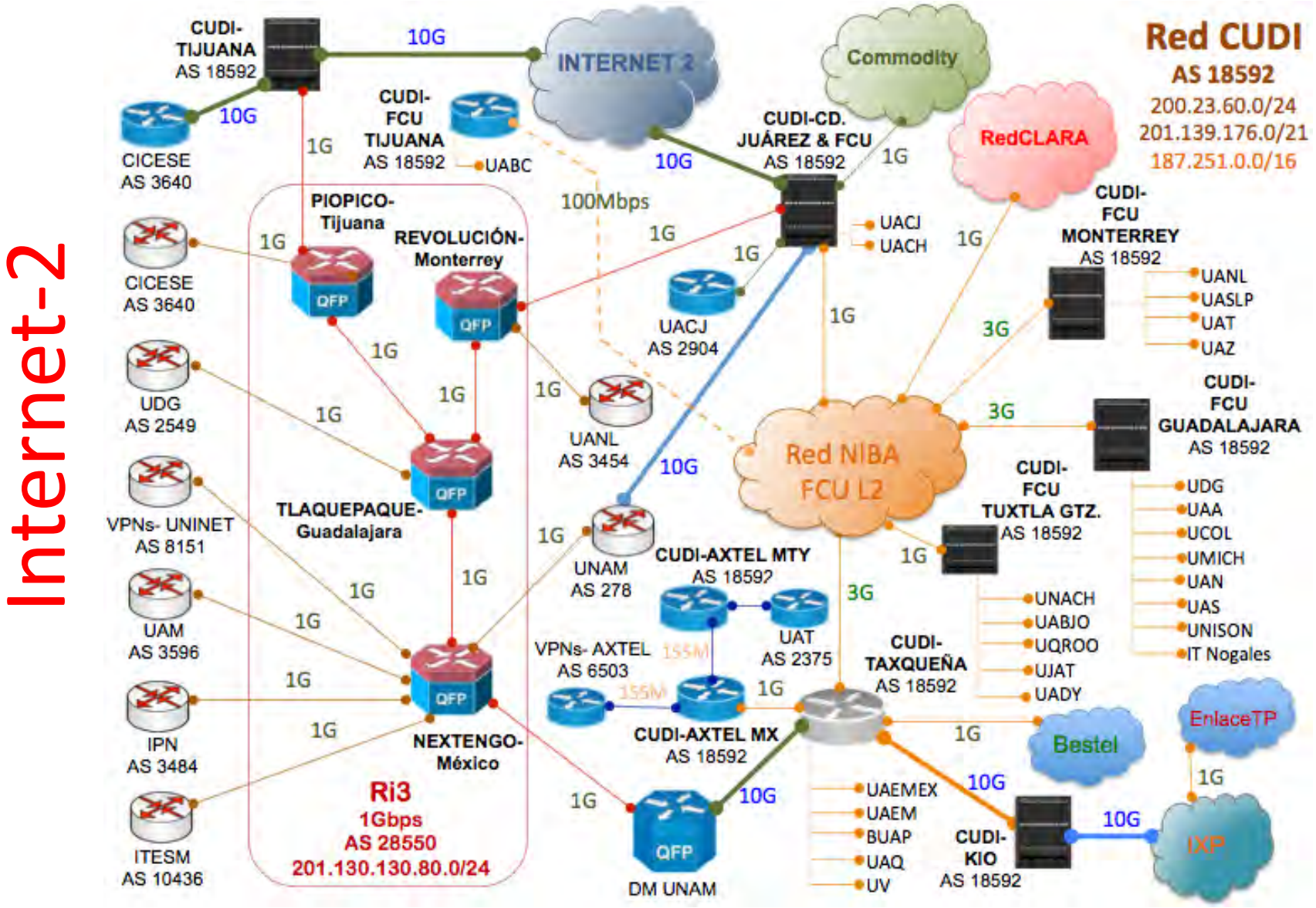
Max Mflop/s: 15565.32

Avg Mflop/s: 14335.02

	Sirio	Berenice	Bakliz	KanBalam	Miztli
marca	CRAY	SGI	HP	HP	HP
procesador	Vectorial	R10000	Alpha EV67	Opteron Dual Core	Intel E2670 8 cores
número de procesadores	4	40	32	1,368	5,312 8334
rendimiento numérico (GFlops)	1.02	15.6	80	7,113	118,000 228,000
memoria (Gigabytes)	0.512	10	32	3,016	23,000 45,000
almacenamiento (Gigabytes)	19	170	1,000	160,000	750,000

Internet-2


Backbone de la red CUDI
| Ver mapa interactivo |
RedNIBA
10G Dorsal RedNIBA
Proyectado


Internet-2
Red CUDI
AS 18592
200.23.60.0/24
201.139.176.0/21
187.251.0.0/16
CUDI-
TIJUANA
AS 18592
10G
INTERNET 2
Commodity
RedCLARA
CUDI-
FCU
TIJUANA
AS 18592
UABC
CICESE
AS 3640
10G
1G
CUDI-CD.
JUÁREZ & FCU
AS 18592
100Mbps
PIOPICO-
Tijuana
REVOLUCIÓN-
Monterrey
QFP
UACJ
UACH
UACJ
AS 2904
CUDI-
FCU
MONTERREY
AS 18592
UANL
UASLP
UAT
UAZ
3G
CICESE
AS 3640
UDG
AS 2549
UANL
AS 3454
Red NIBA
FCU L2
CUDI-
FCU
GUADALAJARA
AS 18592
UDG
UAA
UCOL
UMICH
UAN
UAS
UNISON
IT Nogales
CUDI-
FCU
TUXTLA GTZ.
AS 18592
TLAQUEPAQUE-
Guadalajara
VPNs- UNINET
AS 8151
UNAM
AS 278
CUDI-AXTEL MTY
AS 18592
UAT
AS 2375
VPNs- AXTEL
AS 6503
155M
UAM
AS 3596
CUDI-
TAXQUEÑA
AS 18592
UNACH
UABJO
UQROO
UJAT
UADY
IPN
AS 3484
NEXTENGO-
México
CUDI-AXTEL MX
AS 18592
Bestel
EnlaceTP
Ri3
1Gbps
AS 28550
201.130.130.80.0/24
ITESM
AS 10436
DM UNAM
UAEMEX
UAEM
BUAP
UAQ
UV
CUDI-
KIO
AS 18592
IXP

NARIT

THE ASTROPHYSICAL JOURNAL

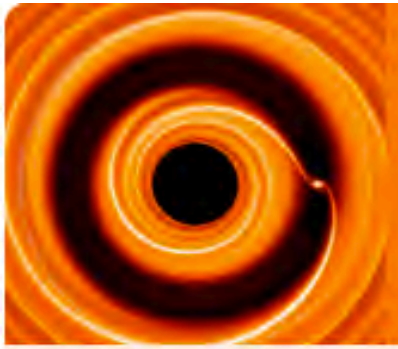
Torques on Low-mass Bodies in Retrograde Orbit in Gaseous Disks

F. J. Sánchez-Salcedo[1], Raúl O. Chametla[2], and A. Santillán[3]

Published 2018 June 20 • © 2018. The American Astronomical Society. All rights reserved.

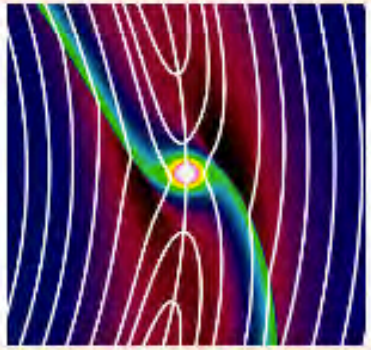
The Astrophysical Journal, Volume 860, Number 2

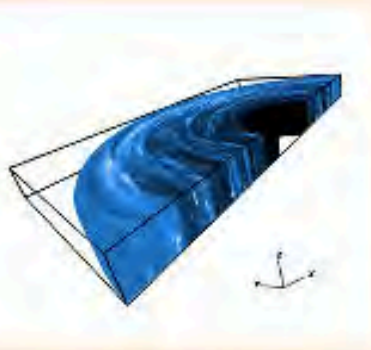
ALMA
Proxima Centauri
Guillem Anglada 2017
4 AU



FARGO3D

A versatile HD/MHD code that runs on clusters of CPUs or GPUs, with special emphasis on protoplanetary disks.





FARGO3D is the successor of the FARGO code, which you can still find in the legacy part of this site. The main features of FARGO3D are:

- Cartesian, cylindrical or spherical geometry
- 1-, 2- or 3-dimensional calculations
- Orbital advection (aka FARGO) for HD and MHD calculations
- As in FARGO, a simple Runge-Kutta N-body solver may be used to describe the orbital evolution of embedded point-like objects
- No need to know CUDA: you can develop new functions in C and have them translated to CUDA automatically to run on GPUs

FARGO3D was written by Pablo Benítez Llambay (main developper) and by Frédéric Masset.

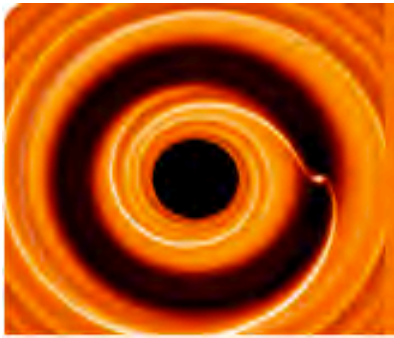
FARGO3D

Features
Download
Documentation
Publications
Acknowledgments

Other

Legacy archive
External links

HPC



FARGO3D

- Solves the equations of hydrodynamics (continuity, Navier-Stokes and energy) and magnetohydrodynamics (MHD) on an Eulerian mesh.
- Multidimensional (1D, 2D & 3D).
- Several geometries (Cartesian, cylindrical and spherical).
- Non inertial reference frames (including shearing box for Cartesian setups).
- Adiabatic or Isothermal Equation of State (EOS). It is easy to implement any other EOS.
- Designed mainly for disks, but works well for general problems.
- Includes ideal MHD (Method Of Characteristics & Constrained Transport).
- Includes magnetic field diffusion (resistivity module).
- Includes the full viscous stress tensor in the three geometries.
- Simple N-body integrator, for embedded planets.
- FARGO algorithm implemented in Cartesian, cylindrical and spherical coordinates.
- The FARGO or "orbital advection" scheme is also implemented for MHD.
- Possible runtime visualization.
- Low memory footprint.
- Multiplatform:
    - Sequential Mode, one process on a CPU.
    - Parallel Mode, for clusters of CPU (distributed memory, with MPI).
    - One GPU (CUDA without MPI).
    - Parallel GPU Mode, for clusters of GPUs (mixed MPI-CUDA version).

Also in this section

Features
Download
Documentation
Publications
Acknowledgments

HPC

Here are quoted a few speed up ratios between GFARGO, and the standard, double precision FARGO code on one CPU core. These speed up ratios are obtained in the limit of a large mesh. In the case of the comparison problem of de Val-Borro et al. (2006), the small size of the mesh (384*128) yields somewhat smaller ratios. The CPU code, for these benchmarks, is always compiled with `gcc`, with the `-O3` and `-ffast-math` options.

Graphics card	CPU	ratio	Platform
8600M GT	Intel(R) Core(TM) 2 Duo@2.4 GHz	9.5x	MacBook Pro (2007)
Quadro FX5800	Opteron AMD 2380@2.5 GHz	90x	Clusters at CEA
Tesla C1070	Intel(R) Xeon(TM) 5570@2.93GHz	30x	Cluster at CEA
GeForce GTX 285	Intel(R) Core(TM)2 Duo E6750@2.66GHz	61x	Linux workstation

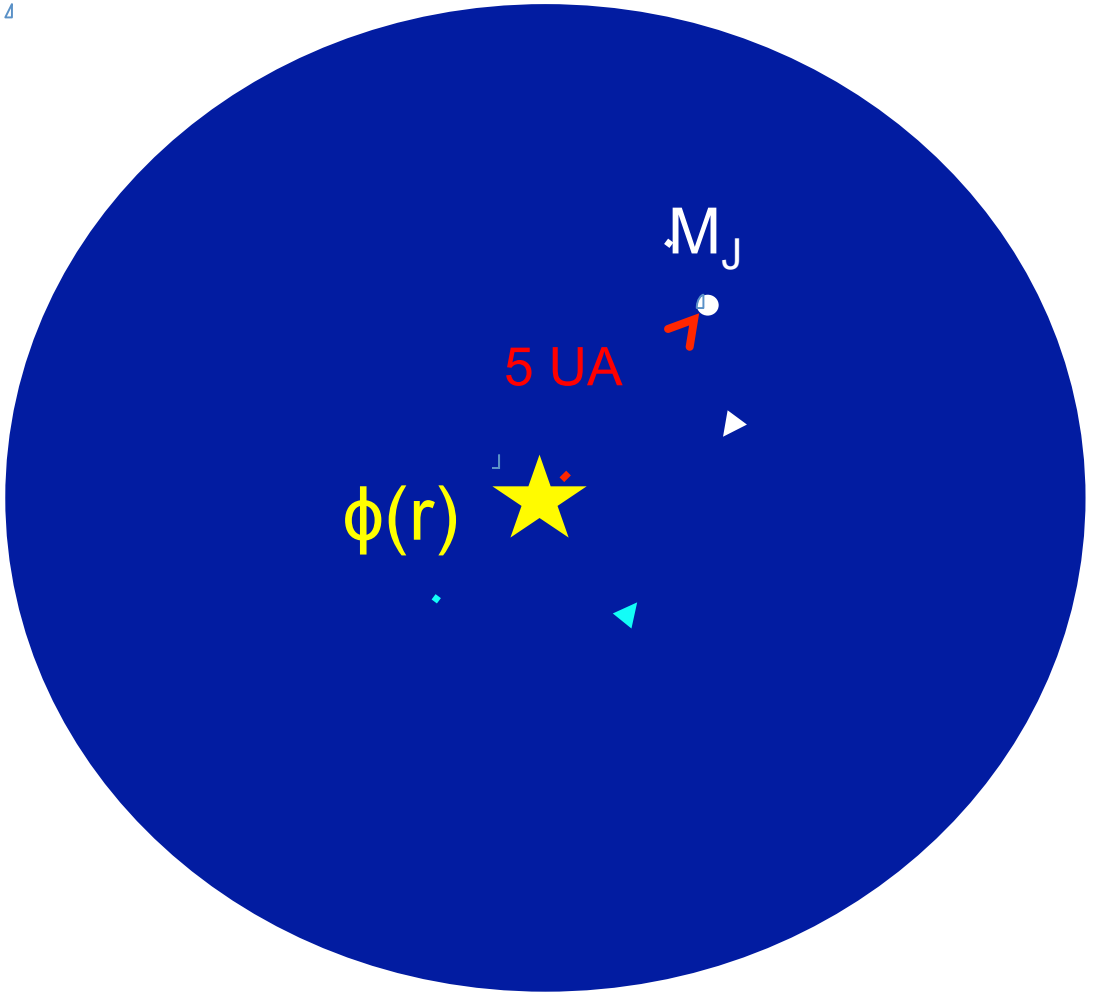
Legacy archive
Acknowledgements
Code Features
Download
User's manual
Specific versions
- FARGO-ADSG
- FARGO-2D1D
- **GFARGO**
- FARGO-AD

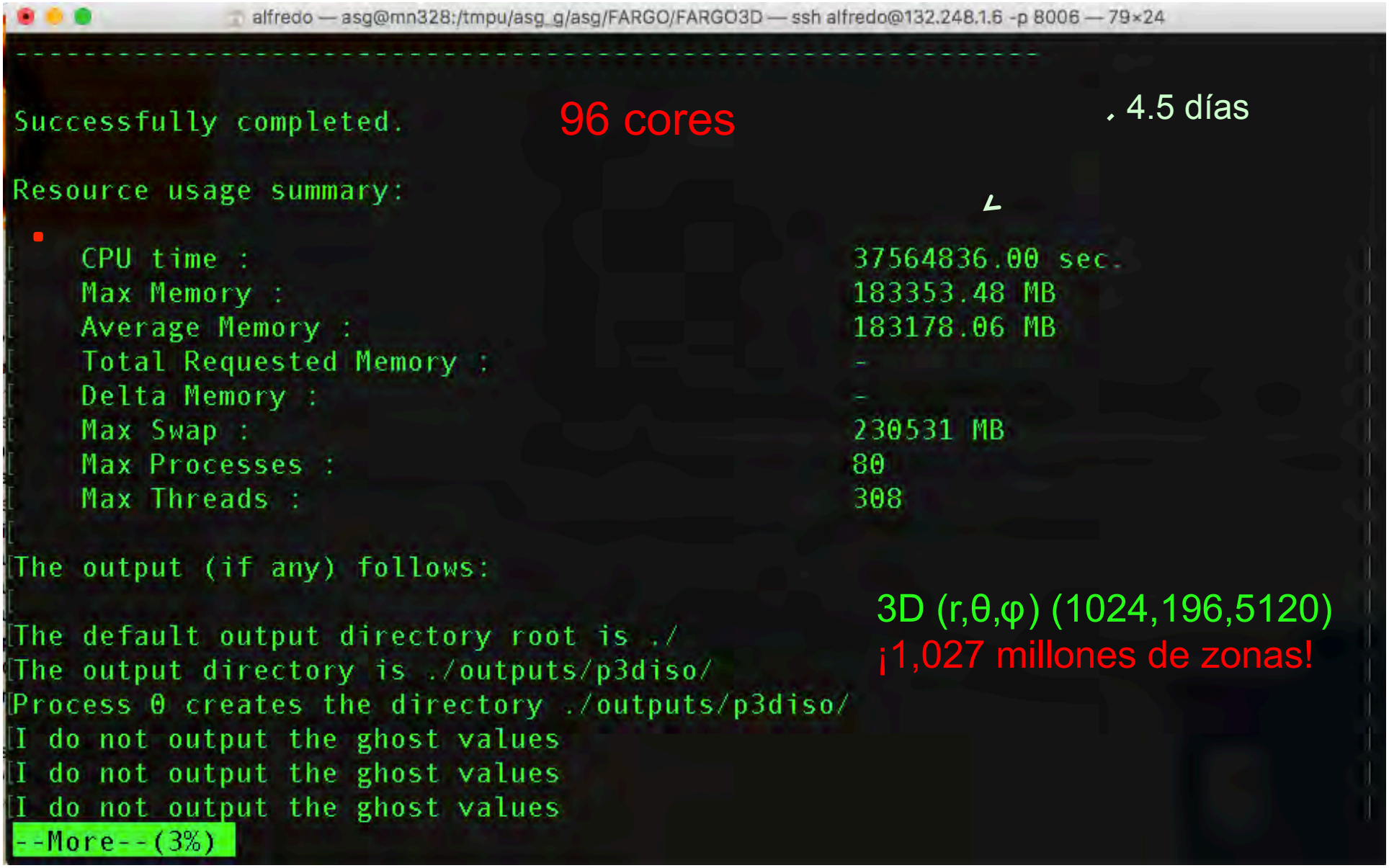
Visualization
Publications

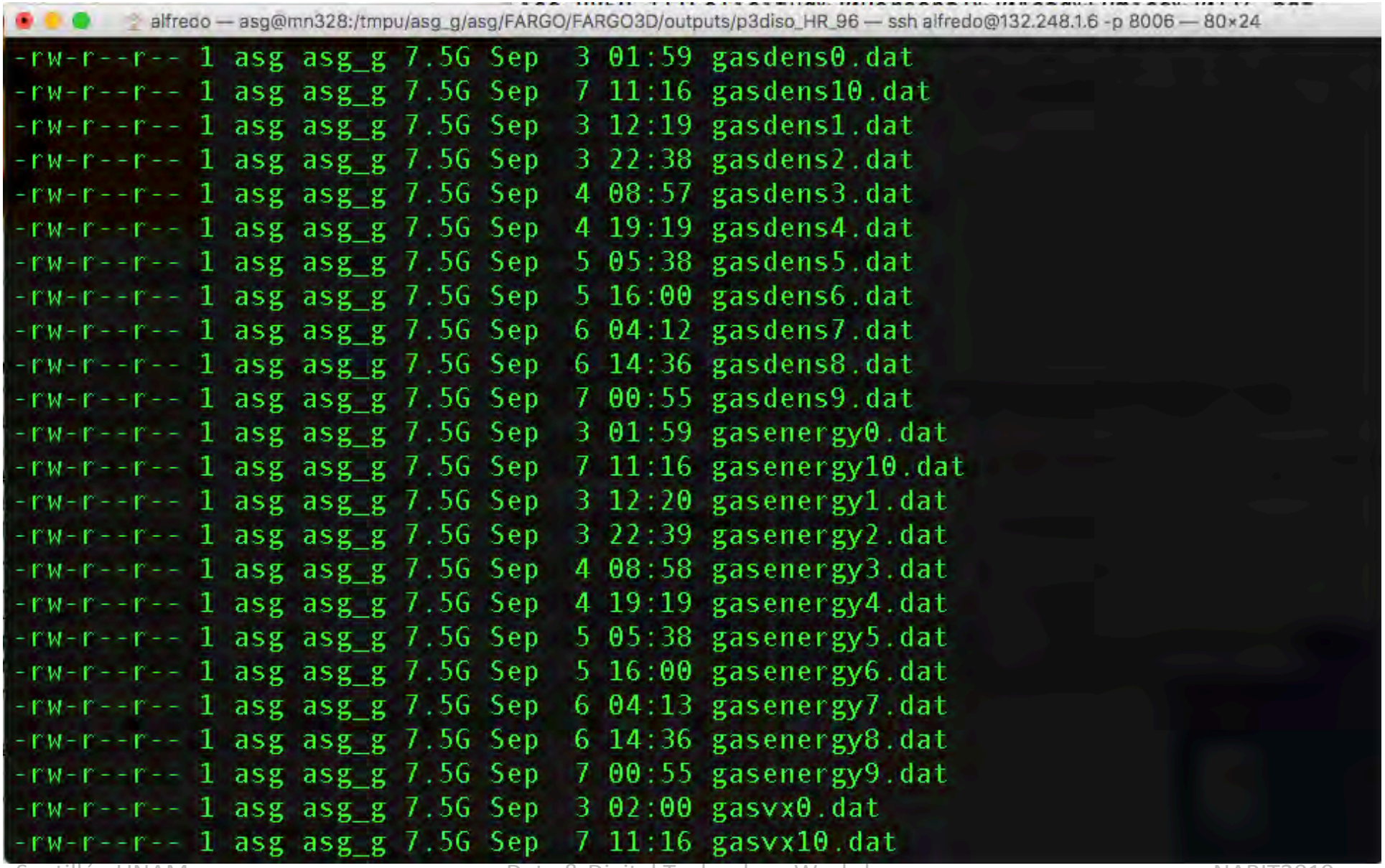
Also in this section
Overview
Download
FAQ
Interactive commands

Initial Conditions

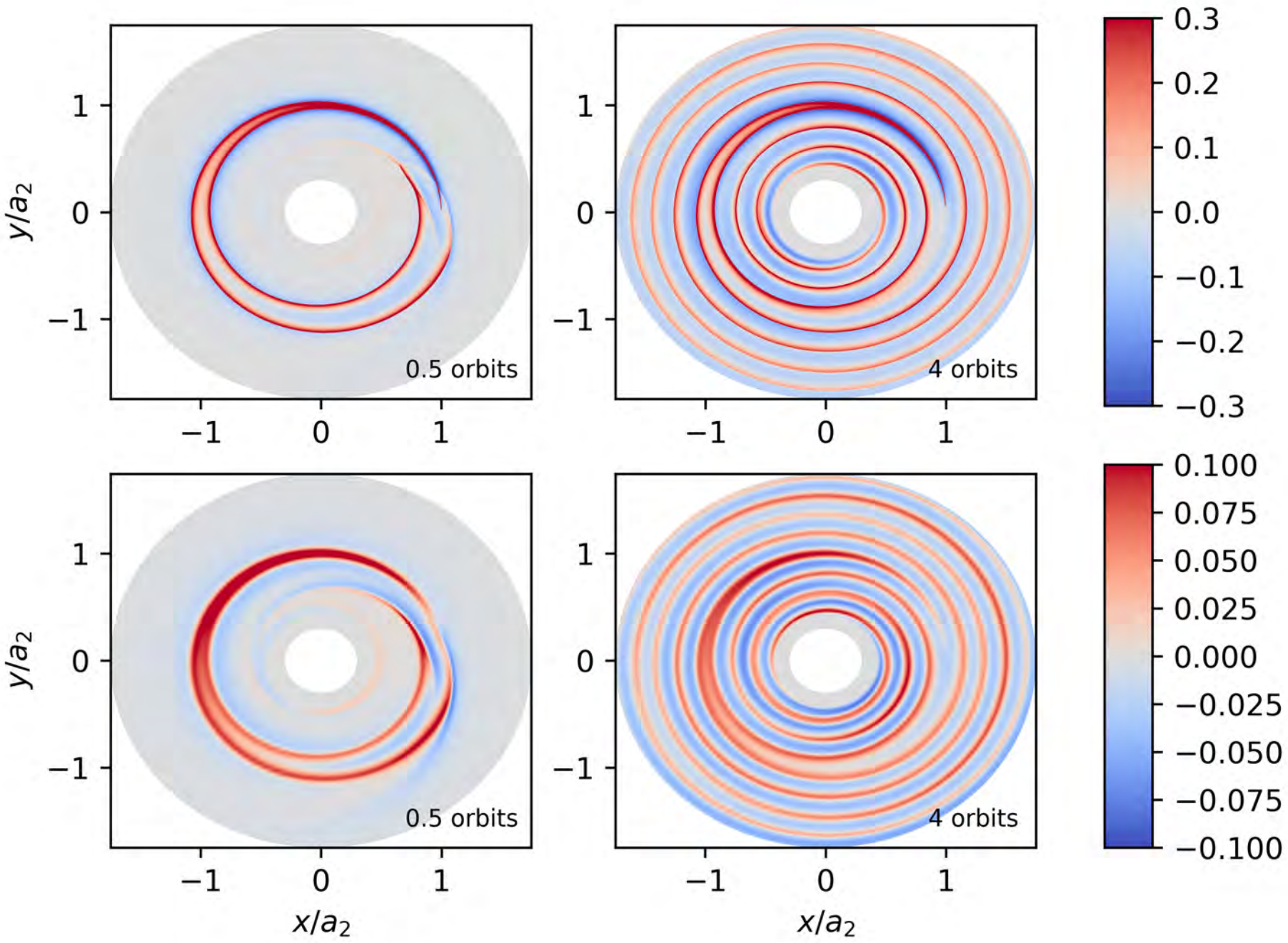
- Keplerian rotation $r^{-1/2}$
- $T = T(r)$
- $M_* = 1\ M_\odot$
- $M_D = 10^{-2}\ M_\odot$
- $M_J = 10^{-3}\ M_\odot$
- $V_J = 13\ km\ s^{-1}$



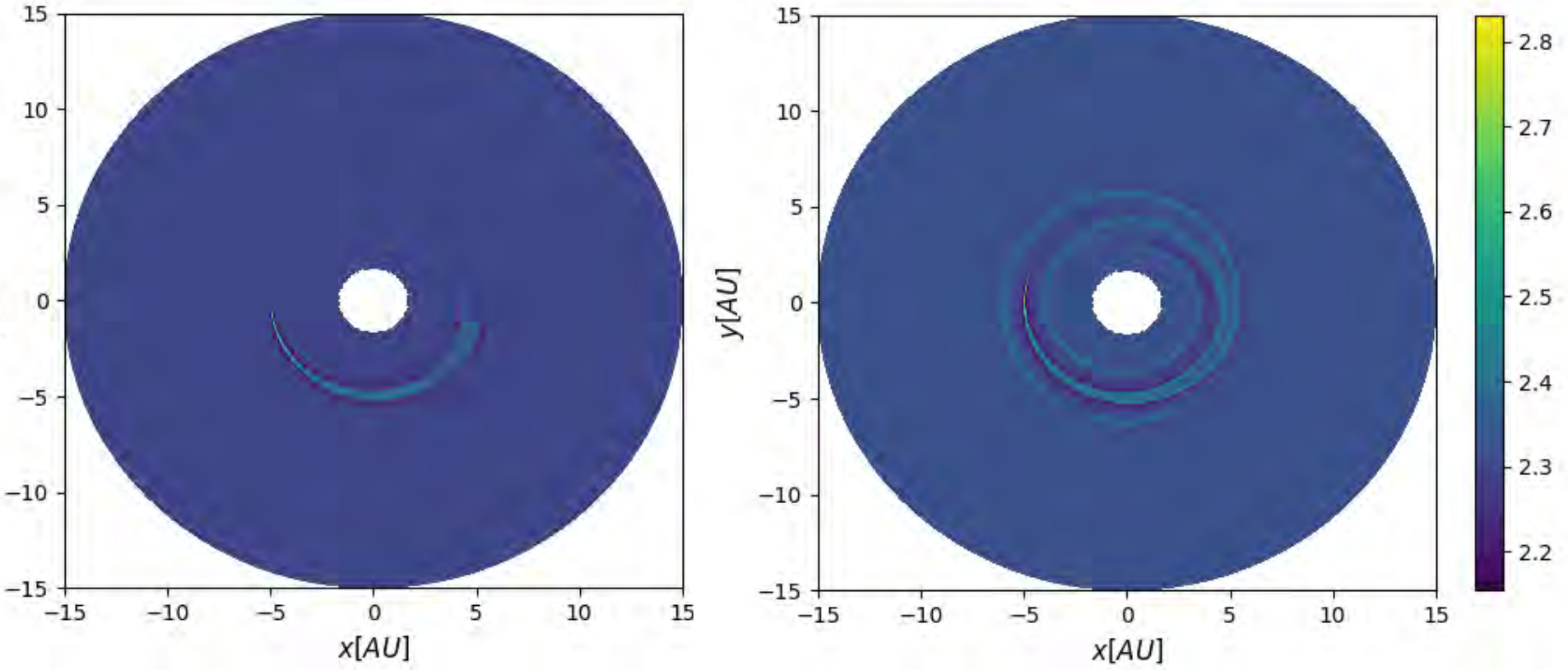

alfredo — asg@mn328:/tmpu/asg_g/asg/FARGO/FARGO3D — ssh alfredo@132.248.1.6 -p 8006 — 79×24
Successfully completed.
96 cores
. 4.5 días
Resource usage summary:
CPU time : 37564836.00 sec.
Max Memory : 183353.48 MB
Average Memory : 183178.06 MB
Total Requested Memory : -
Delta Memory : -
Max Swap : 230531 MB
Max Processes : 80
Max Threads : 308
The output (if any) follows:
3D (r,θ,φ) (1024,196,5120)
¡1,027 millones de zonas!
The default output directory root is ./
The output directory is ./outputs/p3diso/
Process 0 creates the directory ./outputs/p3diso/
I do not output the ghost values
I do not output the ghost values
I do not output the ghost values
--More--(3%)


alfredo — asg@mn328:/tmpu/asg_g/asg/FARGO/FARGO3D/outputs/p3diso_HR_96 — ssh alfredo@132.248.1.6 -p 8006 — 80×24
-rw-r--r-- 1 asg asg_g 7.5G Sep 3 01:59 gasdens0.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 7 11:16 gasdens10.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 3 12:19 gasdens1.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 3 22:38 gasdens2.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 4 08:57 gasdens3.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 4 19:19 gasdens4.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 5 05:38 gasdens5.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 5 16:00 gasdens6.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 6 04:12 gasdens7.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 6 14:36 gasdens8.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 7 00:55 gasdens9.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 3 01:59 gasenergy0.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 7 11:16 gasenergy10.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 3 12:20 gasenergy1.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 3 22:39 gasenergy2.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 4 08:58 gasenergy3.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 4 19:19 gasenergy4.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 5 05:38 gasenergy5.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 5 16:00 gasenergy6.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 6 04:13 gasenergy7.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 6 14:36 gasenergy8.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 7 00:55 gasenergy9.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 3 02:00 gasvx0.dat
-rw-r--r-- 1 asg asg_g 7.5G Sep 7 11:16 gasvx10.dat

alfredo — asg@mn328:/tmpu/asg_g/asg/FARGO/FARGO3D/outputs/p3diso_HR_96 — ssh alfredo@132.248.1.6 -p 8006 — 80×24
[asg@mn328 p3diso_HR_96]$ du -h .
9.1M ./FG000008
9.1M ./FG000000
9.1M ./FG000005
3.1M ./FG000010
9.1M ./FG000001
9.1M ./FG000004
9.1M ./FG000002
9.1M ./FG000007
9.1M ./FG000009
9.1M ./FG000006
9.1M ./FG000003
413G .
[asg@mn328 p3diso_HR_96]$
413 GB
storage


0.5 orbits
4 orbits
0.5 orbits
4 orbits
y/a_2
y/a_2
x/a_2
x/a_2
1
0
−1
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100

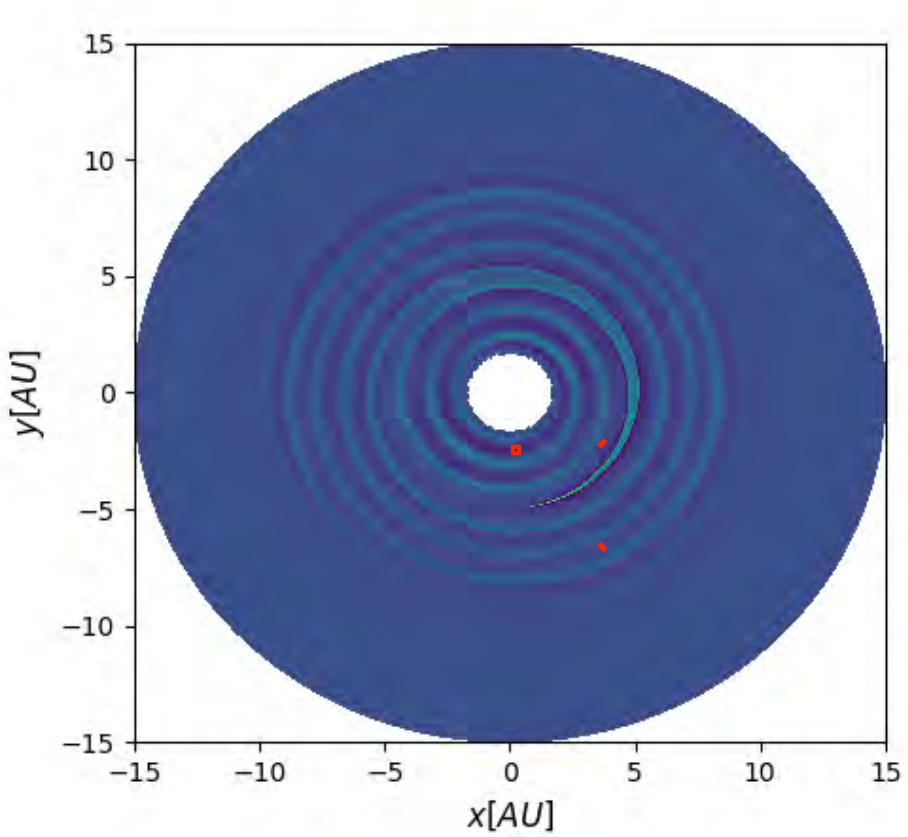
- 2D (r,φ) = (4096,2048)

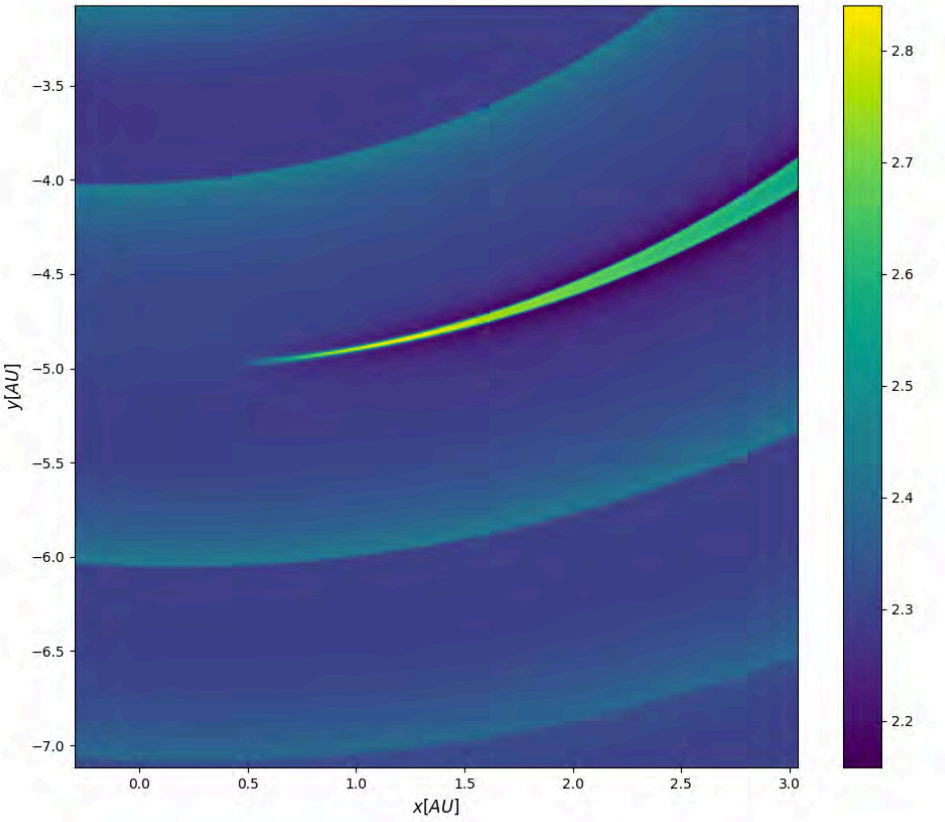


T_1 = ¼ vuelta

T_2 = ¾ vuelta

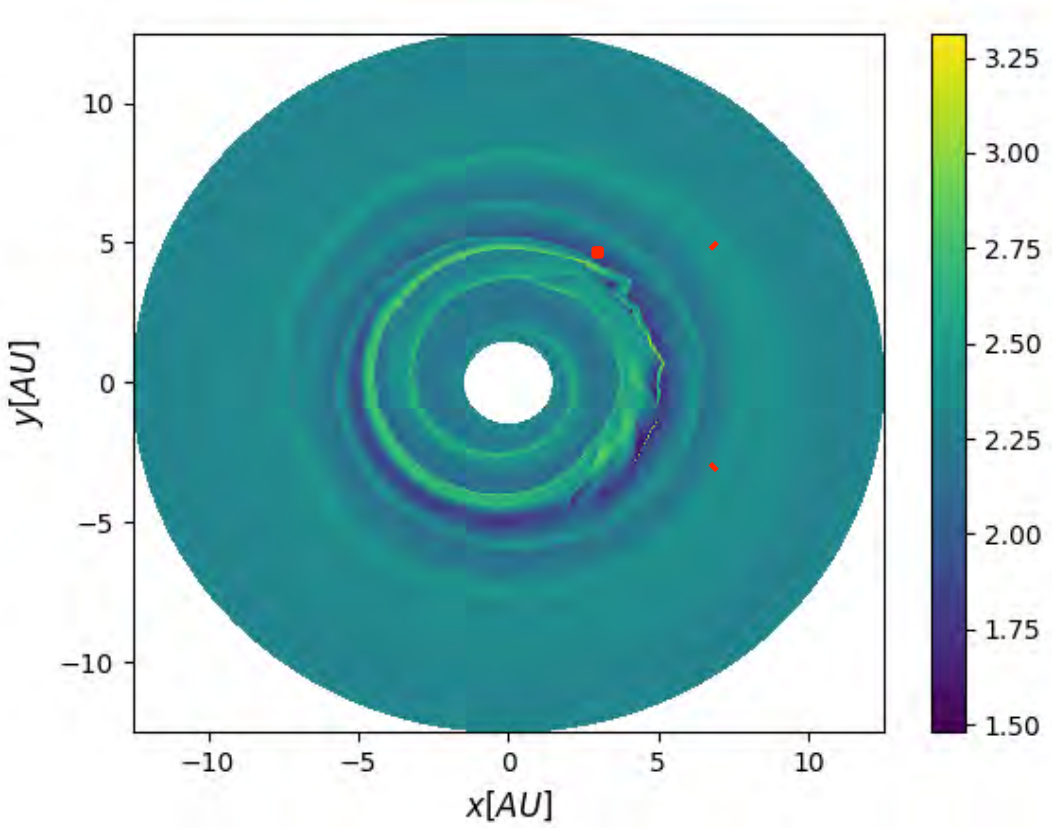
- 2D (r,φ) = (4096,2048)

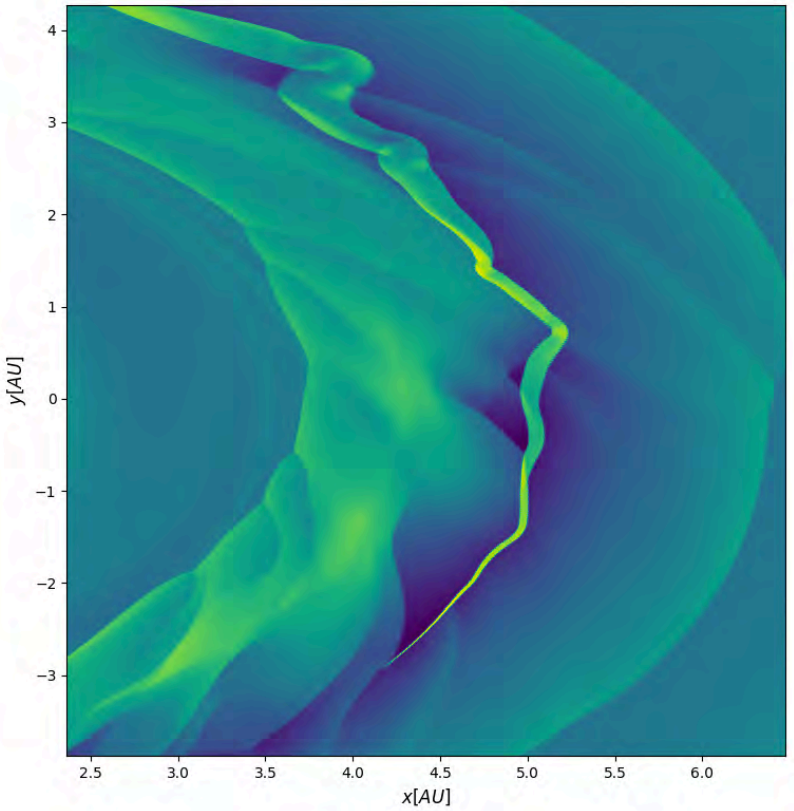


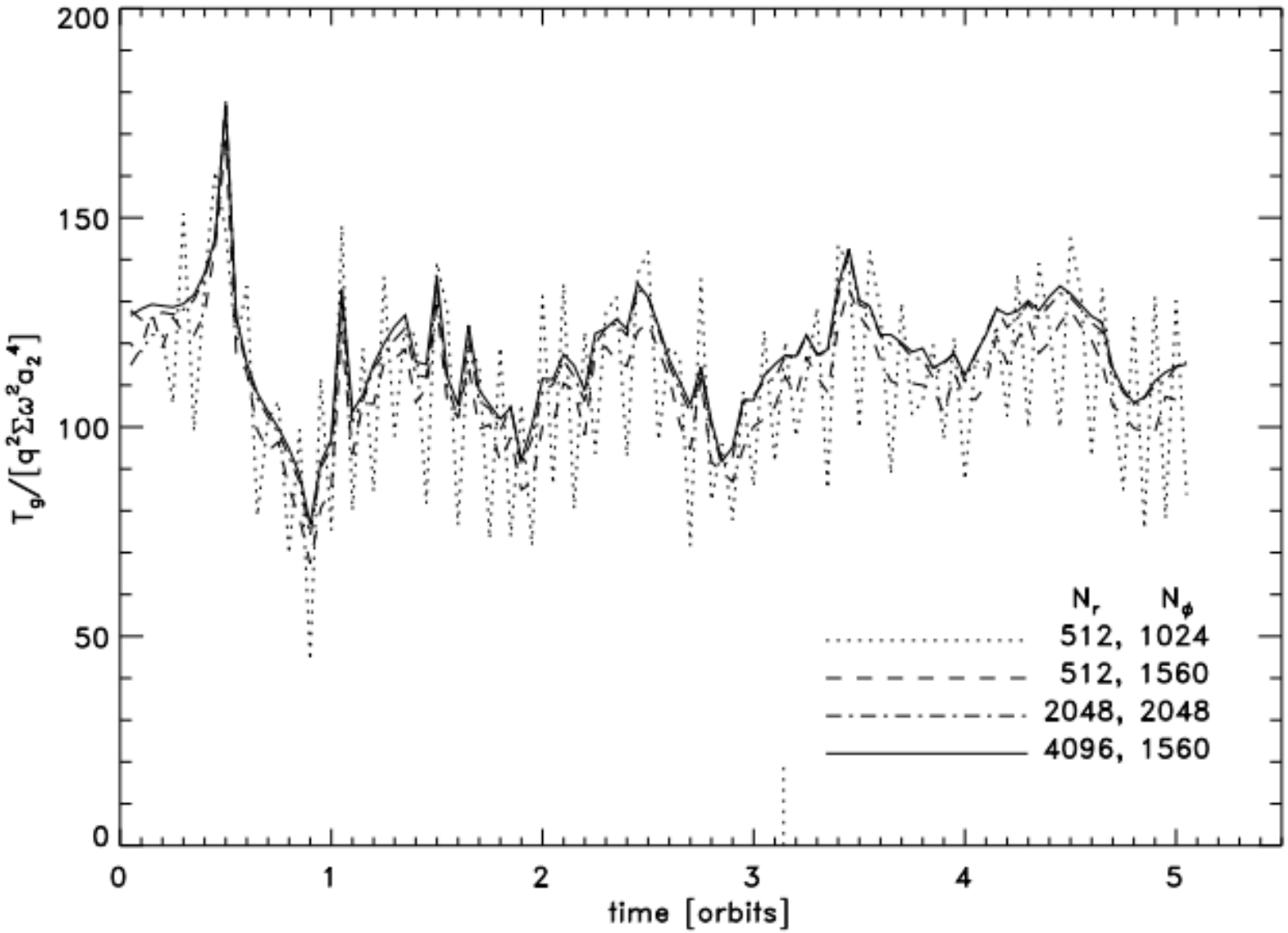


T_3 = 3 vueltas

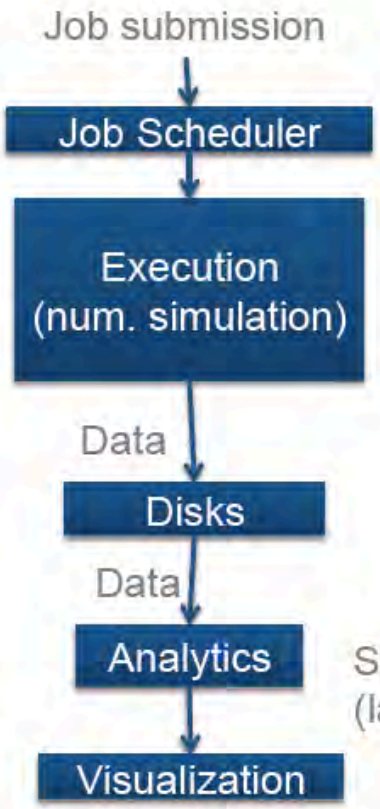
- 2D (r,φ) = (4096,2048)






200
150
100
50
0
0
1
2
3
4
5
time [orbits]
$T_g/[q^2\Sigma\omega^2 a_2^4]$
N_r N_ϕ
512, 1024
512, 1560
2048, 2048
4096, 1560

Traditional Workflow



Not sustainable at Exascale !

Big Machine — Simulation codes may include some analysis

Small Machine (laptop) — Limited support for parallelization

WorkFlow with In-situ Analytics

Big Machine

In-situ analytics:

- **Data reduction**
- **Large scale parallel analytics**
- **On-line monitoring**

Reduced Data Movements

The Data Challenge

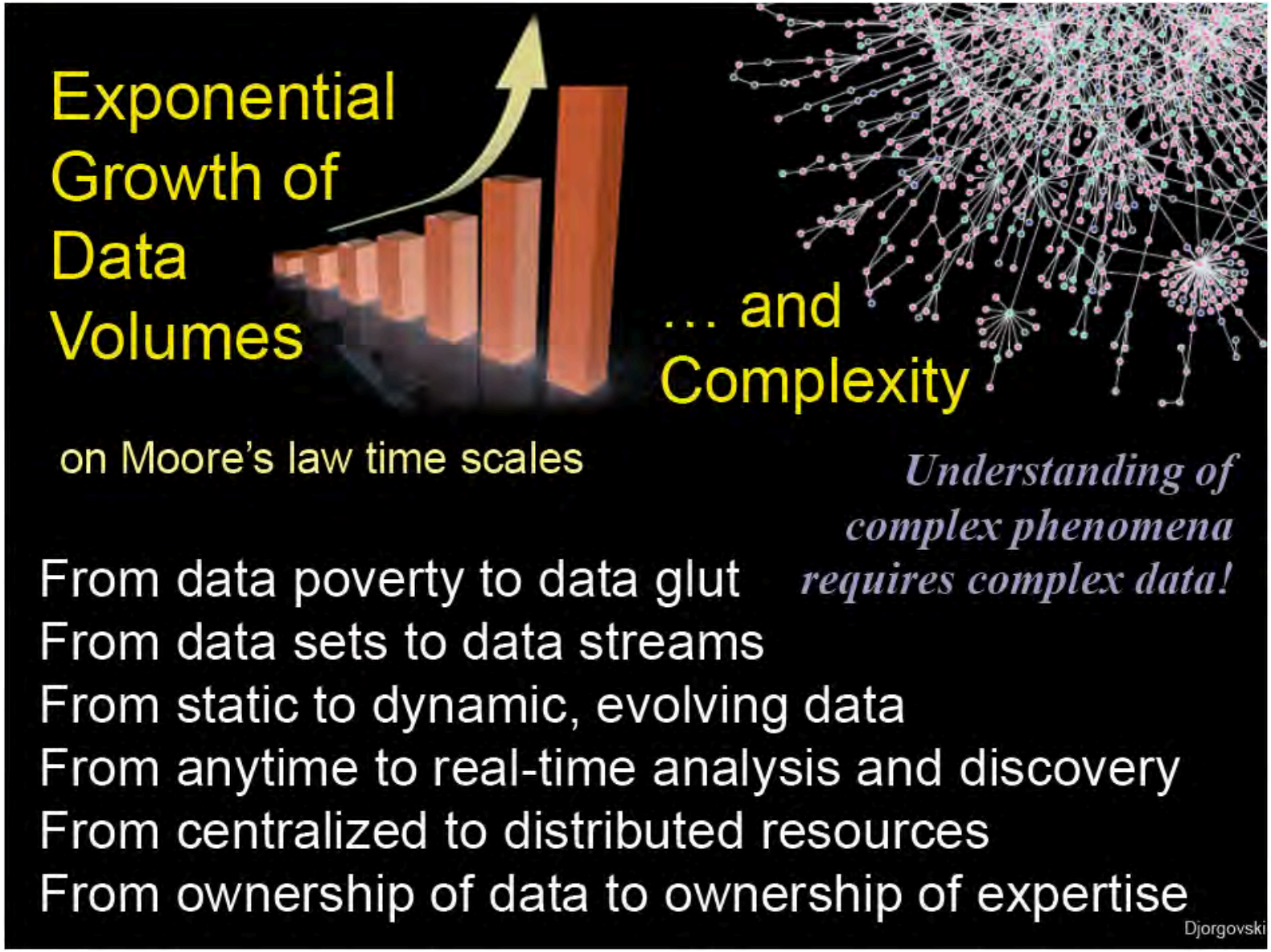
More compute capabilities -> larger simulations -> more data

Usability Challenge:

- How to extract meaningful information from this huge amount of data in a reasonable time
- Analysis tools have not been considered as first class citizen so far. They did not receive the same as simulation codes. Today analysis codes are either:
  - In the simulation codes
  - Scripts (with limited parallelism)
  - Rely on on scientific visualization tools like Paraview/VTK or Visit (reasonable parallelism support)

Performance Challenge:

- Moving data becomes the bottleneck for simulation as well as data analytics
- Compute capabilities increase faster than data transfer ones
- Data movements and storage consume 50%-70% of total energy (ScidacReview 1001)


Exponential Growth of Data Volumes
… and Complexity
on Moore's law time scales
Understanding of complex phenomena requires complex data!
From data poverty to data glut
From data sets to data streams
From static to dynamic, evolving data
From anytime to real-time analysis and discovery
From centralized to distributed resources
From ownership of data to ownership of expertise
Djorgovski