ESnet
ENERGY SCIENCES NETWORK

NDN and Big Data Science

Inder Monga

Interim Director and CTO, ESnet
Interim Director, Scientific Networking Division

Lawrence Berkeley National Lab

Named-Data Network Workshop @ NIST

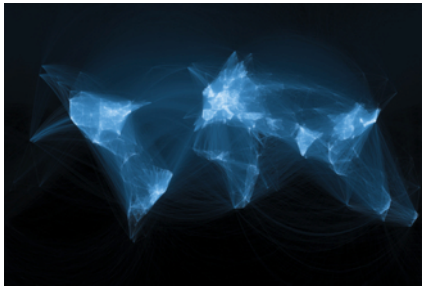
May 31-June 1

U.S. DEPARTMENT OF ENERGY
Office of Science

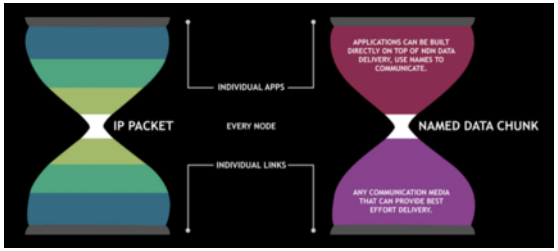
BERKELEY LAB

Agenda

Big Science Data



Global Science Collaborations



NDN for Science

imonga at es dot net

How big is data (visual comparison) *

- Byte
- Kilobyte
- Megabyte
- Gigabyte
- Terabyte
- Petabyte
- Exabyte
- Zettabyte

- One grain of rice
- Cup of rice
- 8 bags of rice
- 3 tractor trailers
- 2 container ships
- Layer of rice over Manhattan
- 2 layers over the United Kingdom
- Fills the Pacific ocean

ESnet

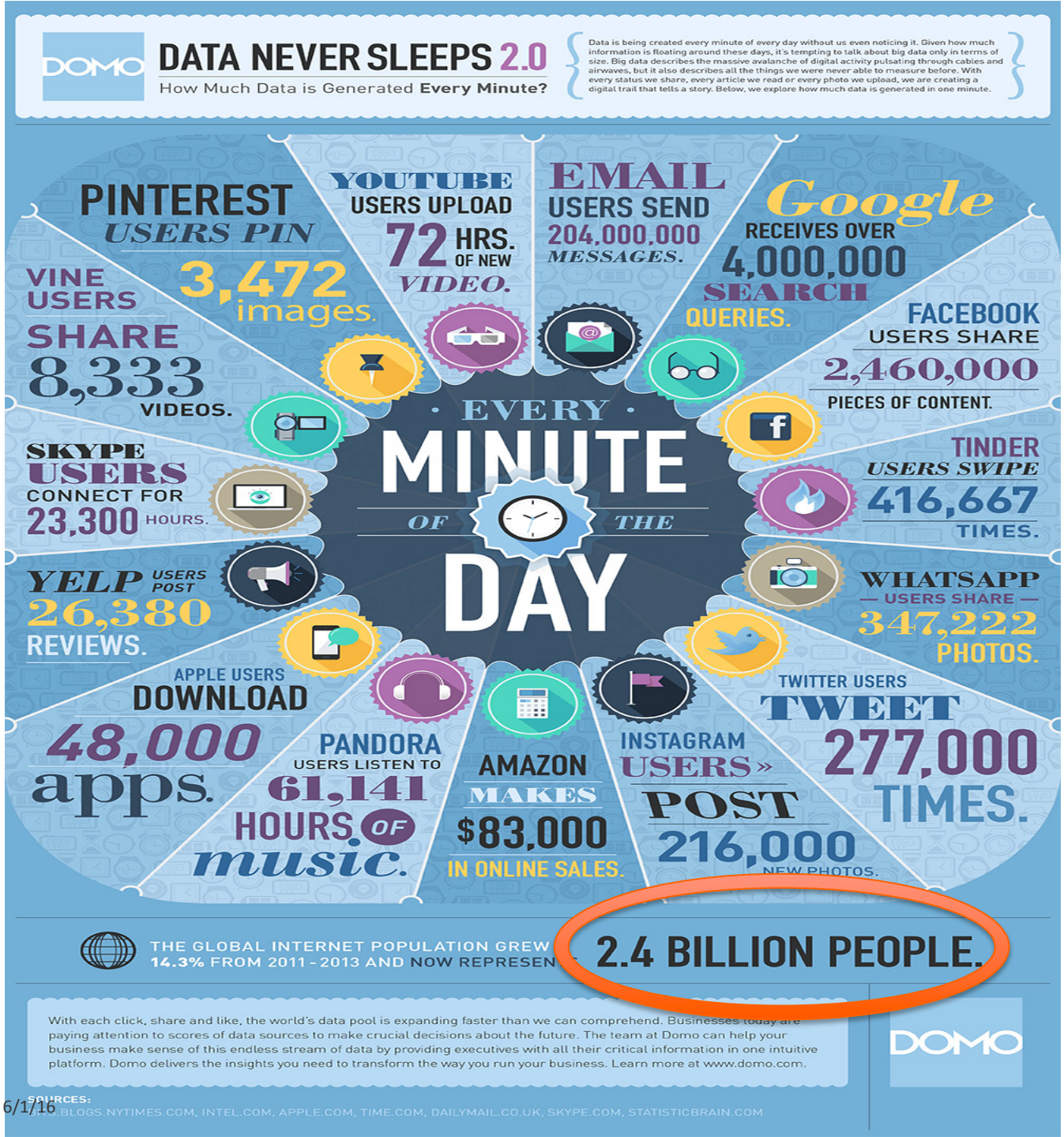
*David Wellman @ Myriad Genetics

Every Instagram photo = 110 KB
216,000 photos are sent to Instagram every minute
This equals 23GB of data per minute

Instagram Data produced per day worldwide = 33 TB
Equal to filling ~1,032 – 32GB iPhones

ESnet

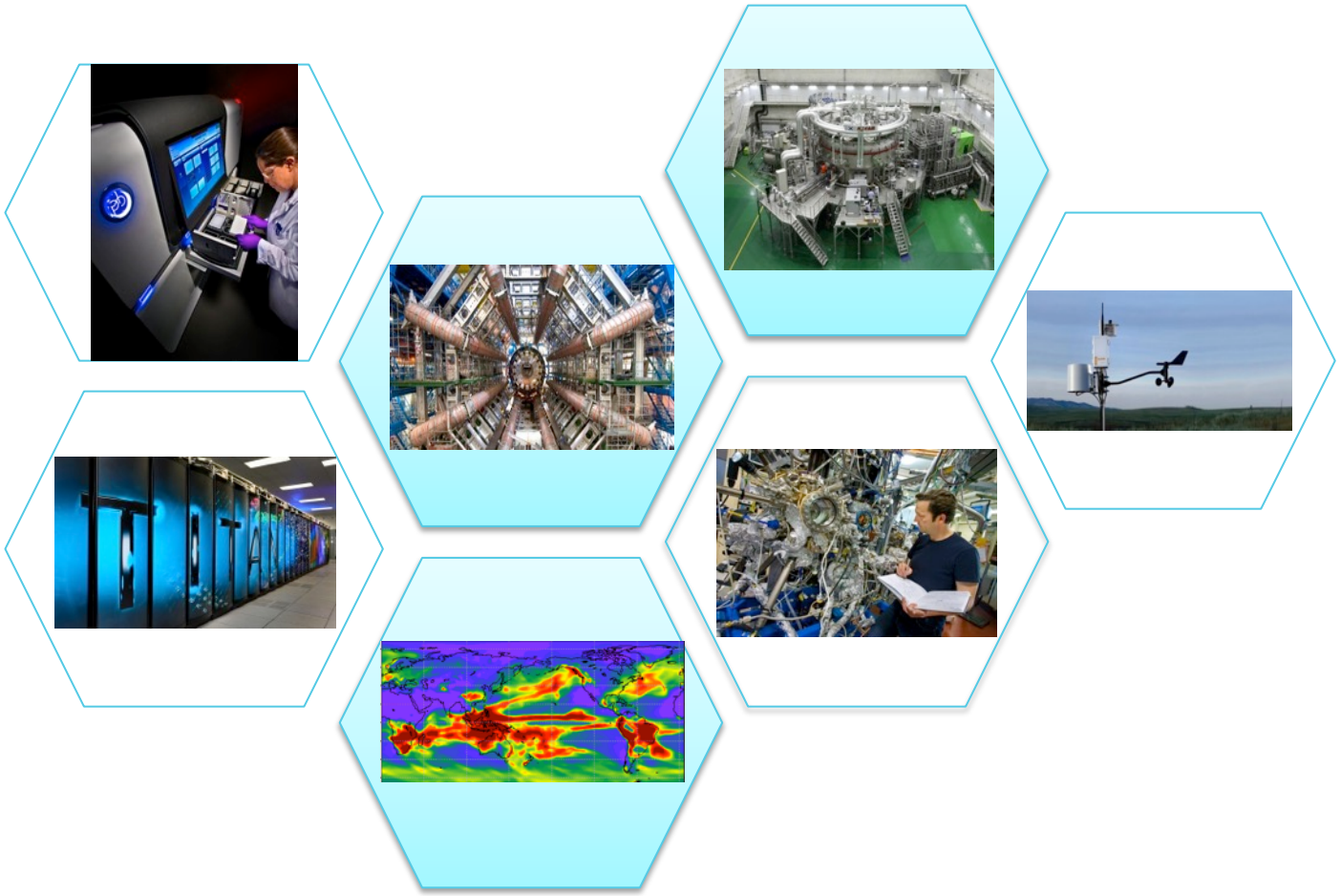
imonga at es dot net



ESnet

imonga at es dot net

DOE Science “Apps”



imonga at es dot net

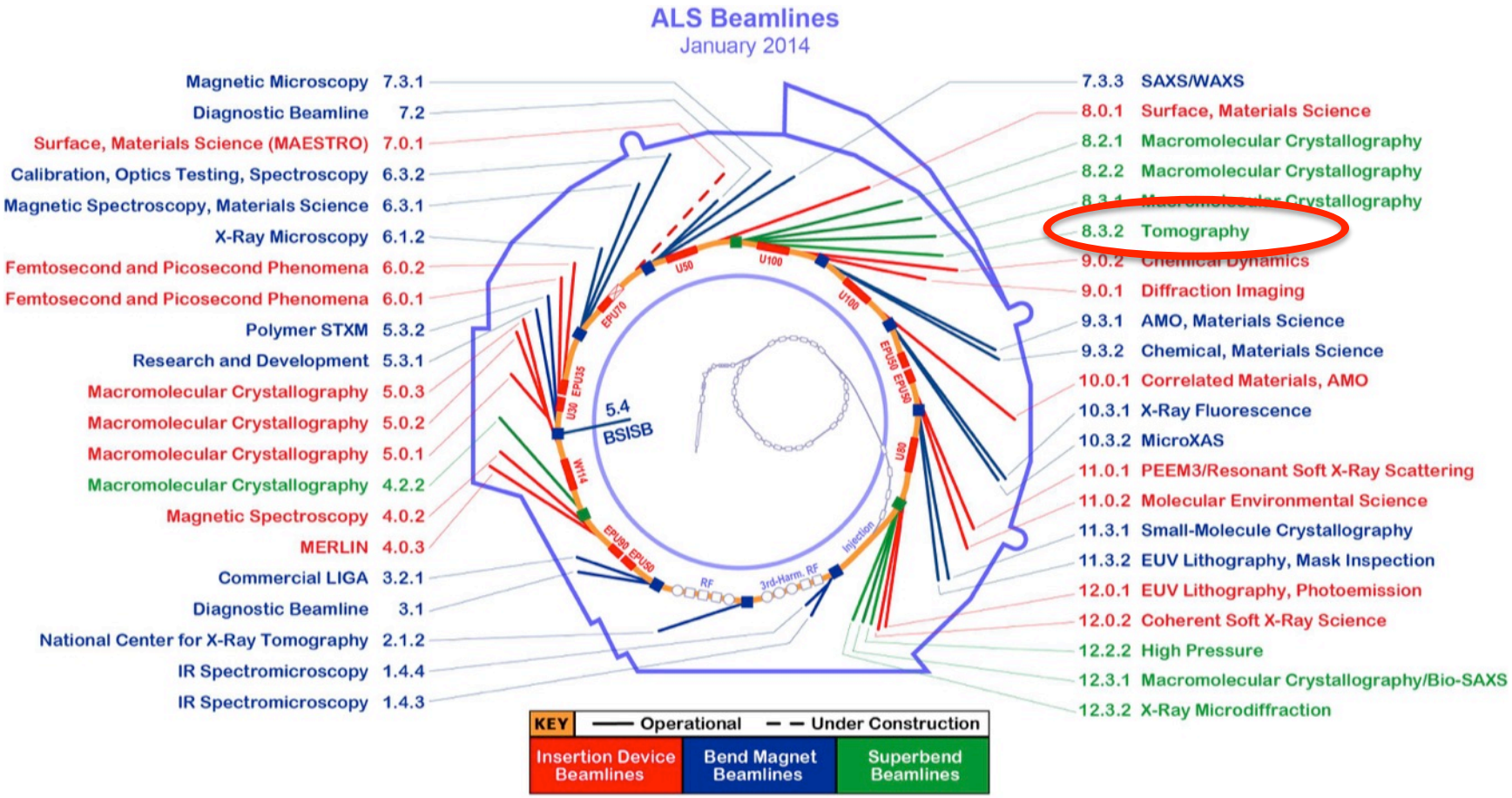
Advanced Light Source

ESnet

Advanced Light Source

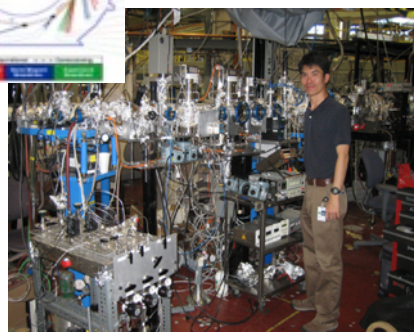
ESnet

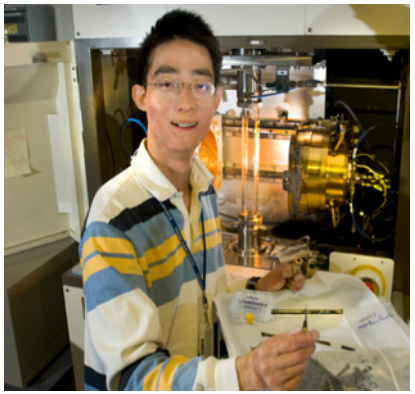
Advanced Light Source

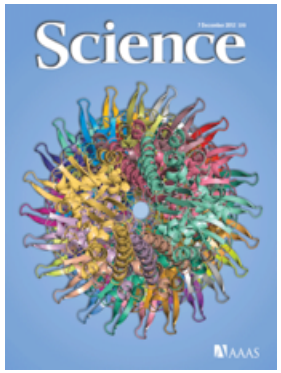


imonga at es dot net

Scenario 1: All too common process of discovery



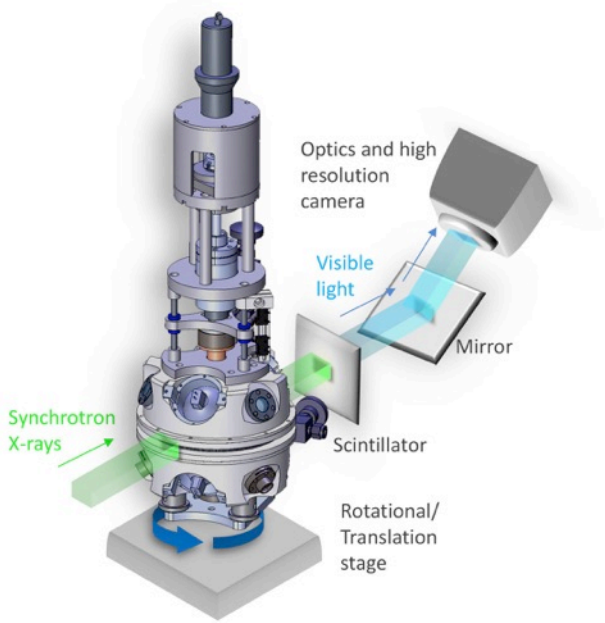


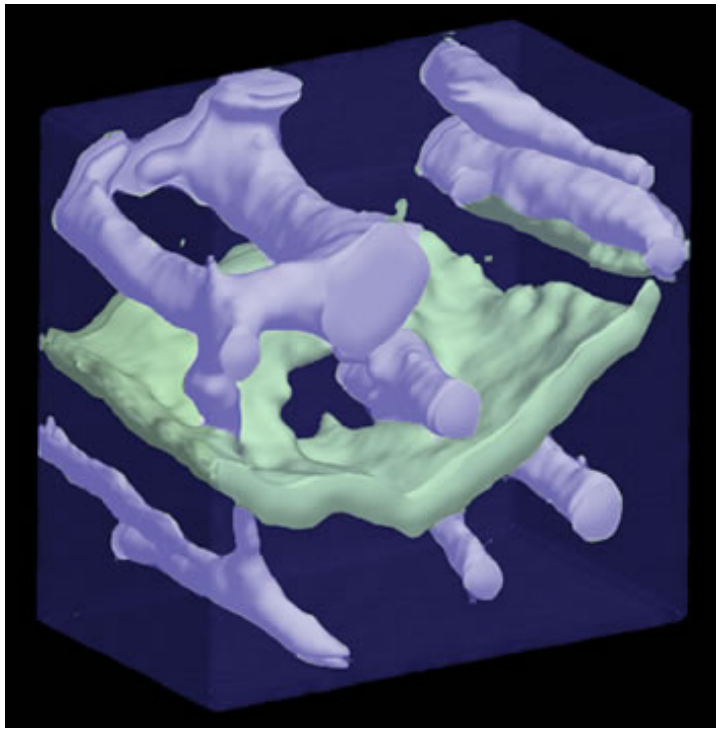


ESnet

imonga at es dot net

Beamline – Capture to Results





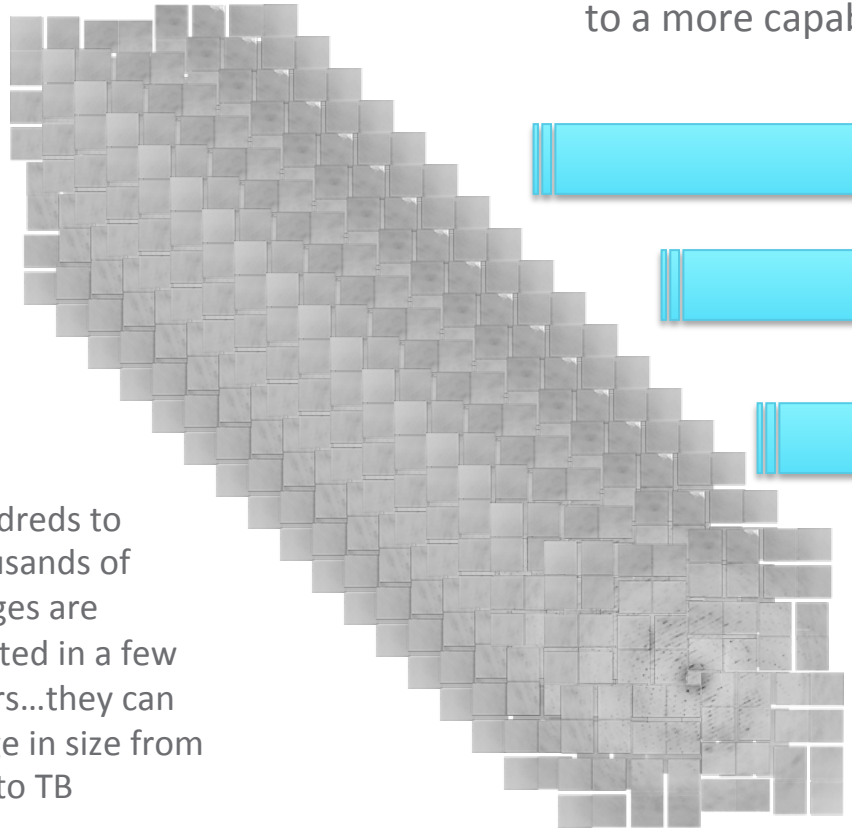
Basic Energy Sciences (BES) supports fundamental research to understand, predict, and ultimately control matter and energy at the electronic, atomic, and molecular levels in order to provide the foundations for new energy technologies and to support DOE missions in energy, environment, and national security.

http://science.energy.gov/bes/

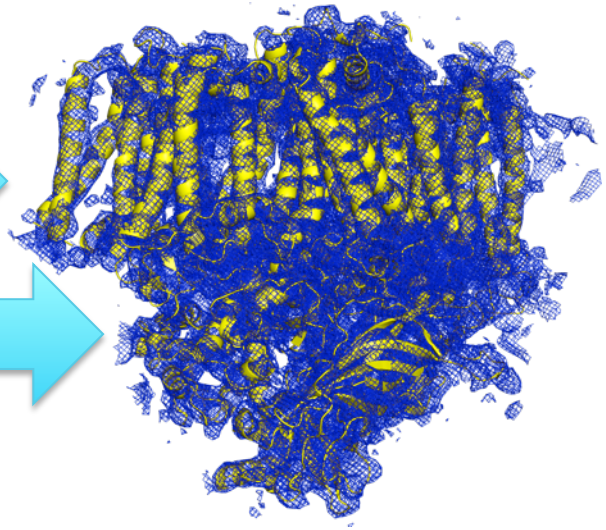
ESnet

Scenario 2: E Pluribus Unum

Processing on this order of magnitude can't be done locally – we need to send (over a network) to a more capable facility



Hundreds to thousands of images are created in a few hours…they can range in size from MB to TB



After processing on a supercomputer, models are created.

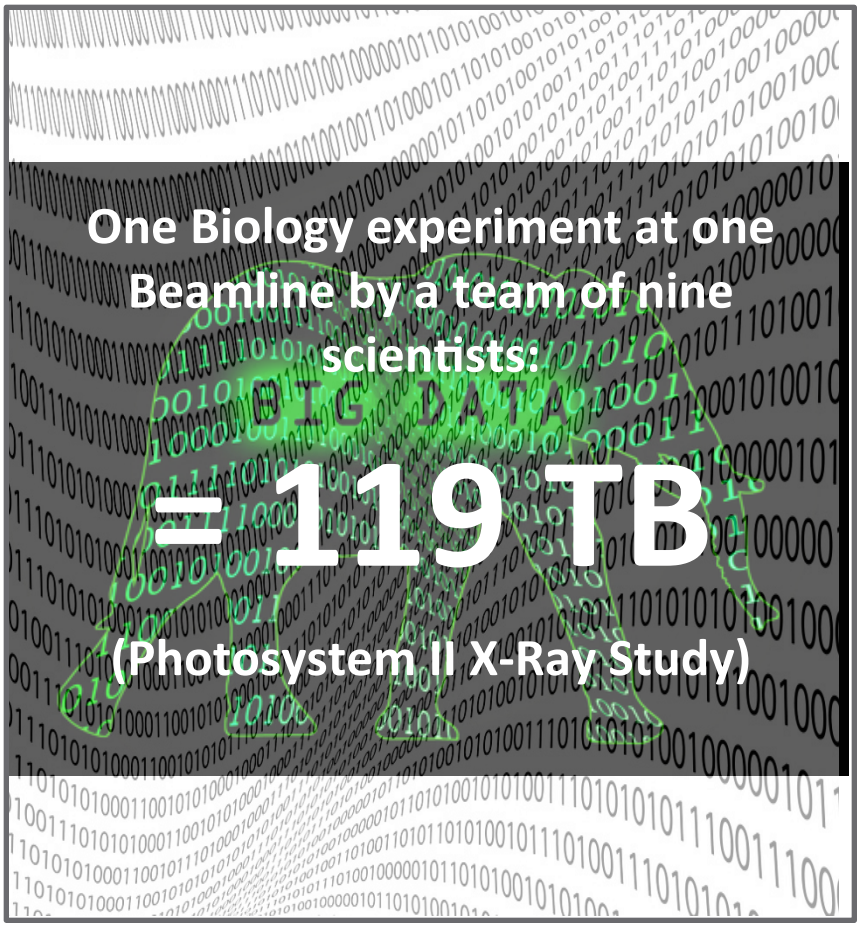
ESnet

Big Data vs. Big Data

Don't Forget:

Instagram Data produced/day worldwide by millions of people

= 33 TB

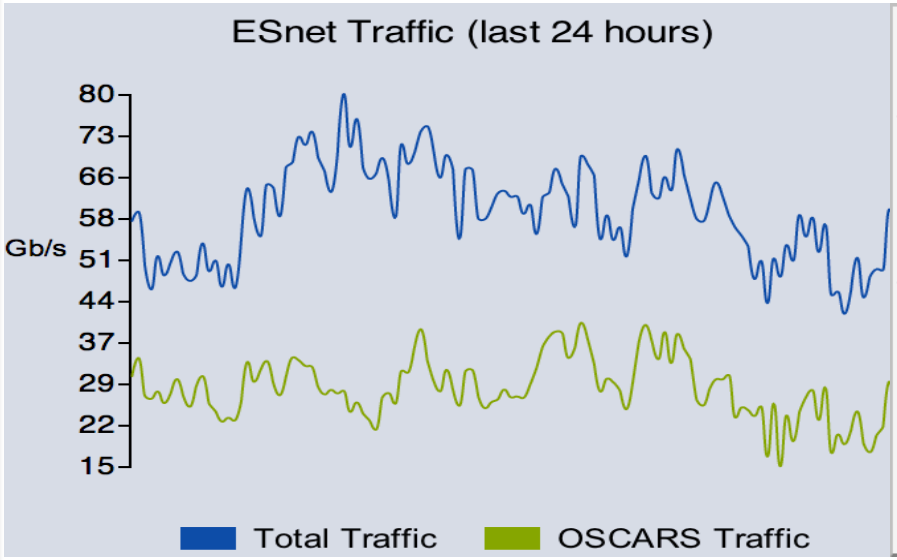


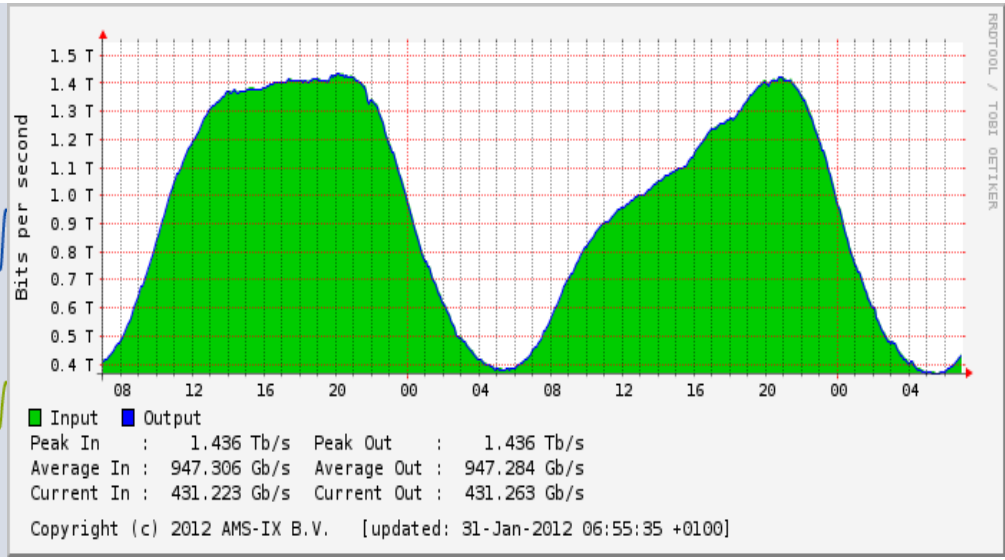
ESnet

Big Science Data in Motion = Elephant Flow!
IoT watching LOL Cats = Mice flow!

ESnet

Elephant Data vs. Mice Data Behavior

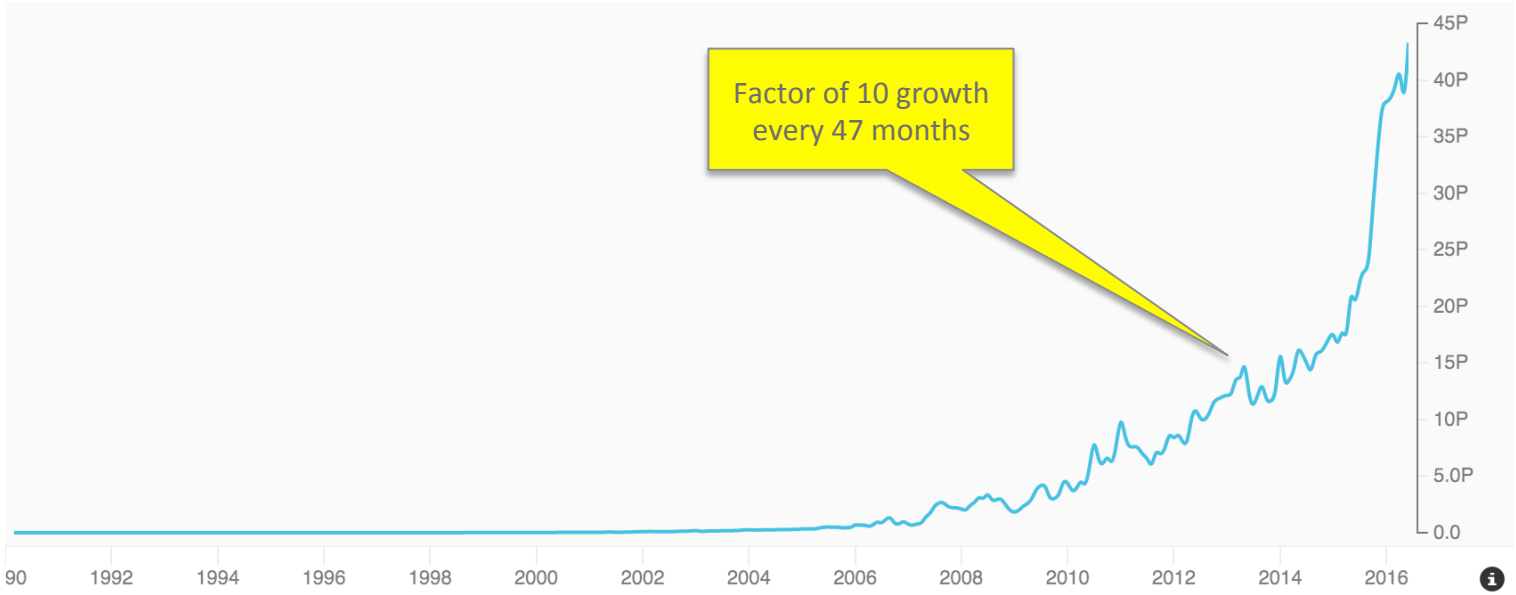




Science Data Transferred Monthly by ESnet

Traffic Volume

Available at https://my.es.net/network/traffic-volume



February 2016

	Bytes	Percent of Total	One Month Change	One Year Change
OSCARS	10.46 PB	25.2%	+0.0147%	+148%
LHCONE	11.22 PB	27.0%	+3.90%	+770%
Normal traffic	19.82 PB	47.8%	+9.49%	+77.7%
Total	41.49 PB		+5.44%	+149%

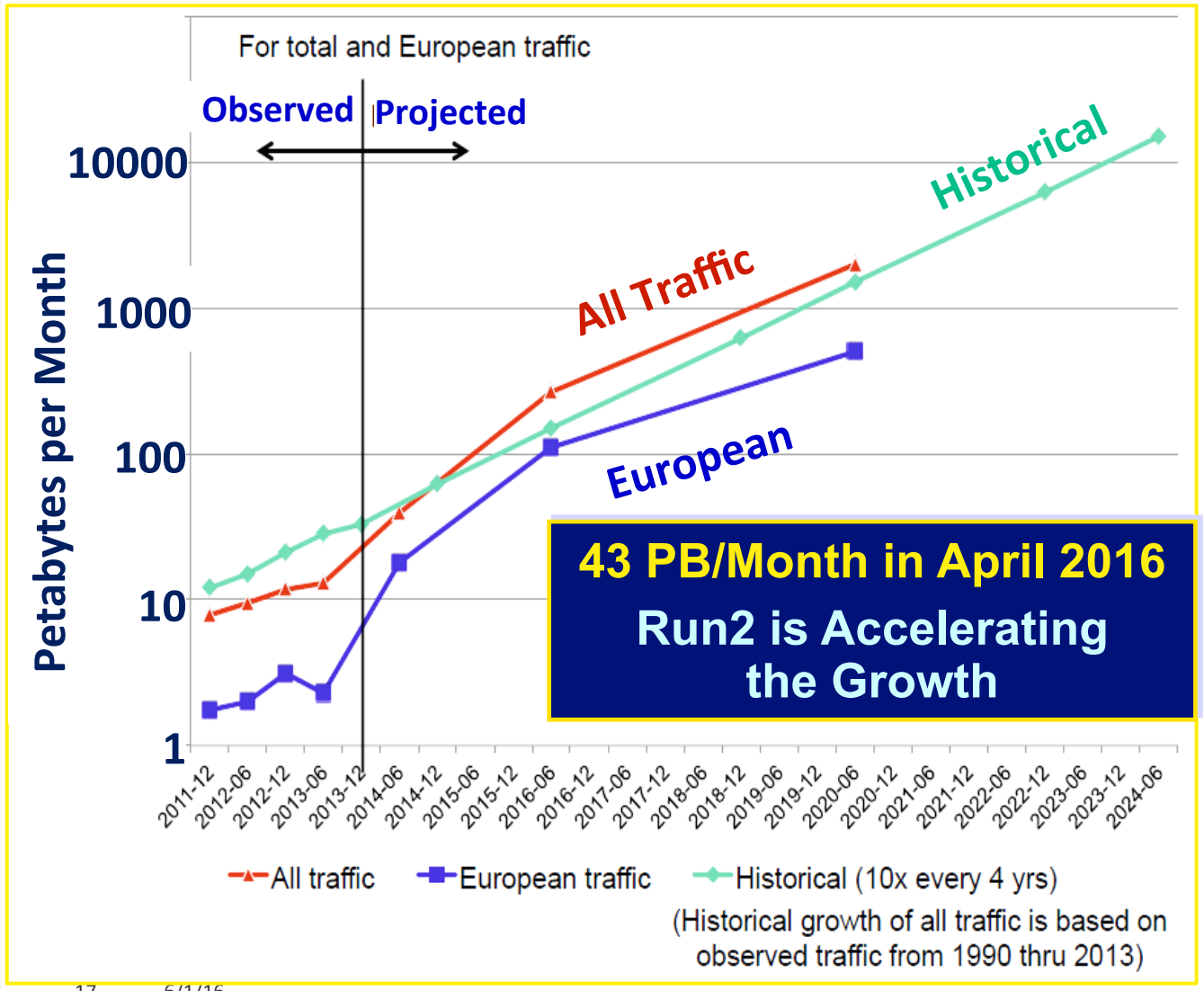
Pt-to-pt circuits

LHCONE (T1-T1/2) traffic

ESnet

imonga at es dot net

Traffic growth at blistering rates



Projected Traffic Reaches

1 Exabyte Per Month. by ~2020

10 EB/Mo. by ~2024

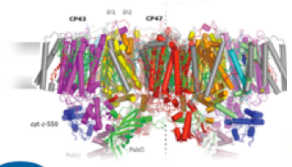
Slide from Harvey Newman

ESnet

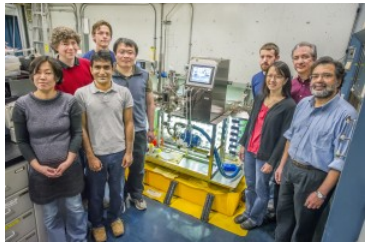
imonga at es dot net

Superfacility: interconnection of multiple facilities via the network

Researchers from Berkeley Lab and SLAC conducted protein crystallography experiments at LCLS to investigate photoexcited states of PSII, with near-real-time computational analysis at NERSC.



LCLS ⟷ ESnet ⟷ NERSC



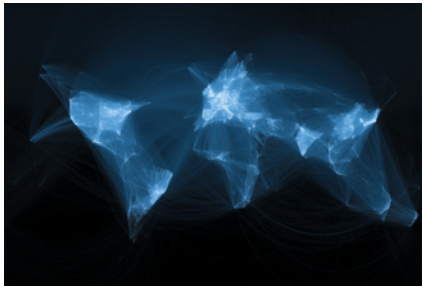
"Taking snapshots of photosynthetic water oxidation using femtosecond X-ray diffraction and spectroscopy," *Nature Communications* 5, 4371 (9 July 2014)

ESnet

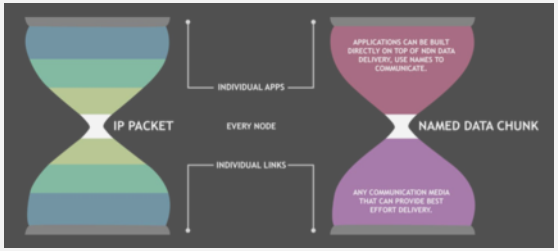
imonga at es dot net

Agenda

Big Science Data



Global Science Collaborations

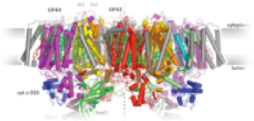


NDN for Science

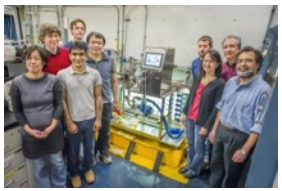
ESnet

Use Case #1

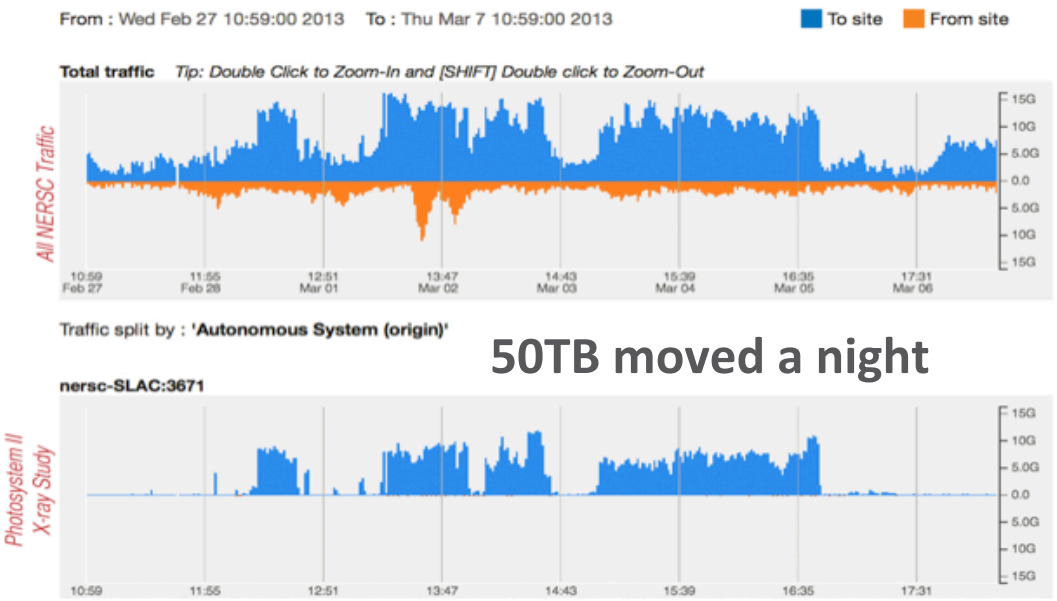
Researchers from Berkeley Lab and SLAC conducted protein crystallography experiments at LCLS to investigate photoexcited states of PSII, with near-real-time computational analysis at NERSC.



LCLS ⟷ ESnet ⟷ NERSC



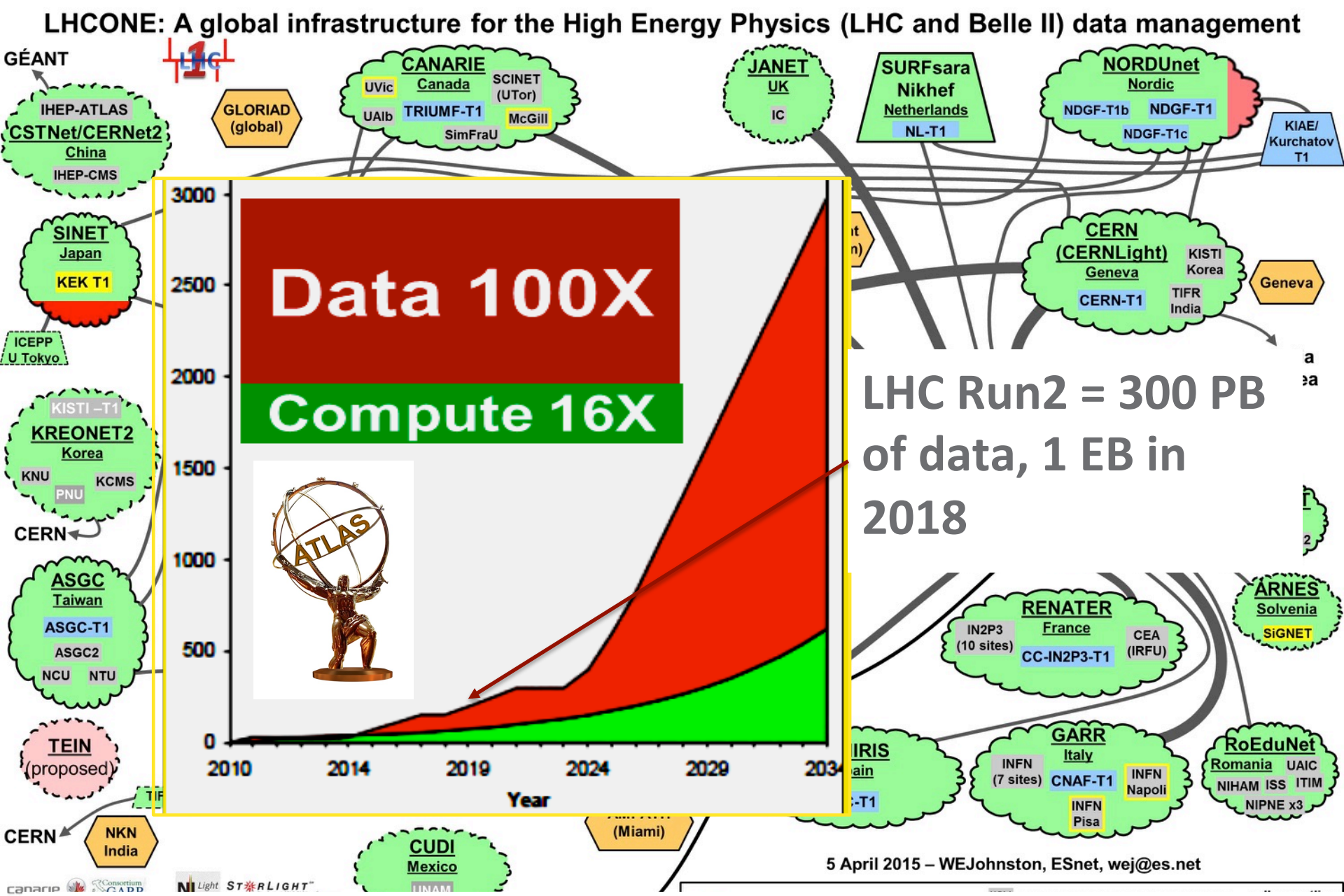
"Taking snapshots of photosynthetic water oxidation using femtosecond X-ray diffraction and spectroscopy," *Nature Communications* 5, 4371 (9 July 2014)



50TB moved a night

ESnet

LHCONE: A global infrastructure for the High Energy Physics (LHC and Belle II) data management



Use Case #2: LHCONE data – multiple replicas, global reach

Use Case 3: Worldwide Earth System Grid Federation Sites

ESnet

Use Case Galore

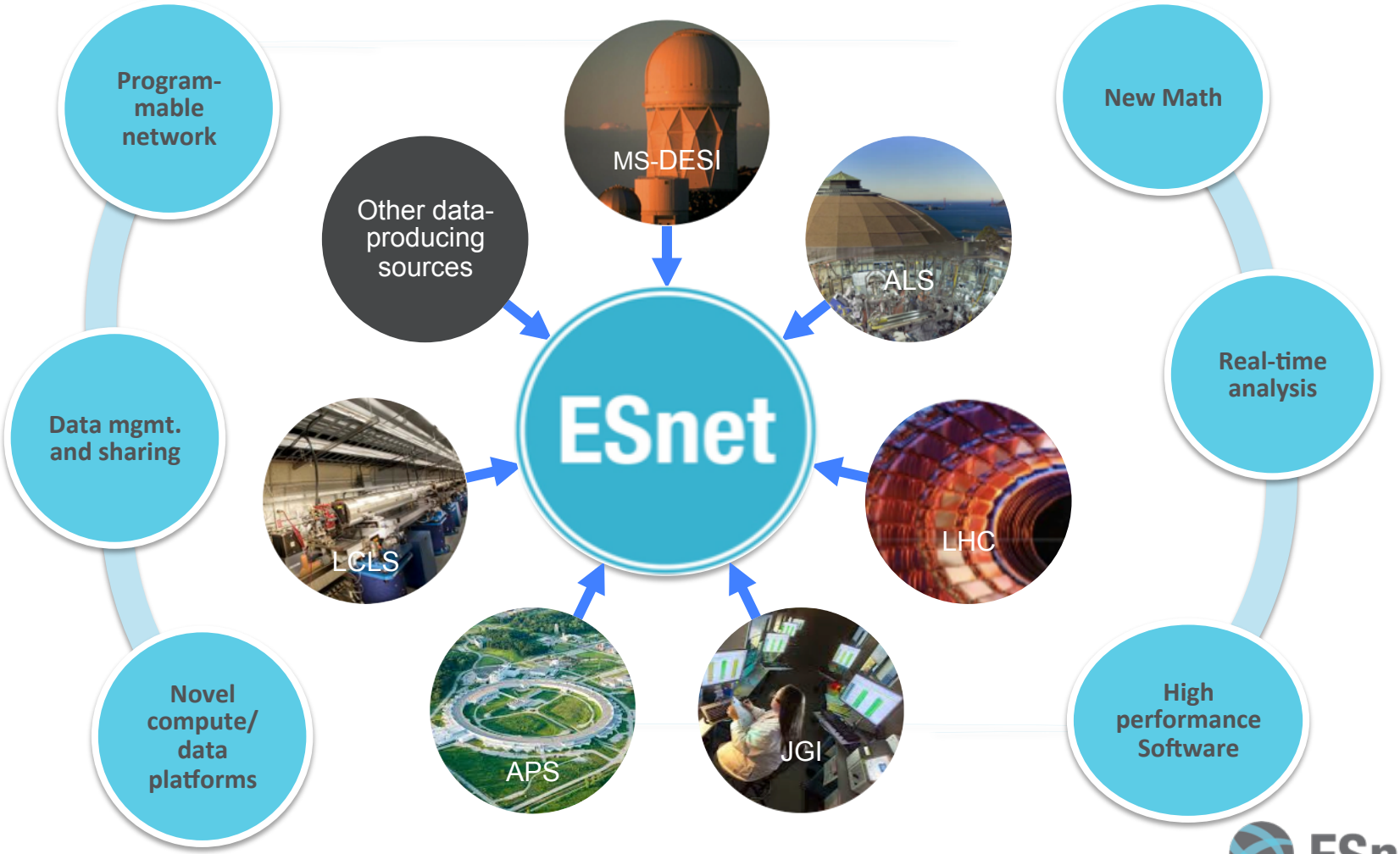
#4 - LCLS: Data coming from Chile, Stored in NCSA, and analyzed among a global collaboration

#5 - SKA: Data coming from South Africa and Australia, analyzed among a global collaboration

#6 - Bio-Health, Precision Medicine, Genomics: Open-data trend, data-sets available at many websites.

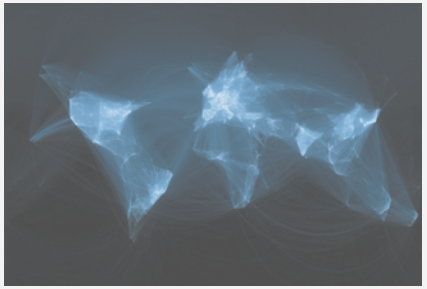
ESnet

Superfacility Vision: A network of connected facilities, software and expertise to enable new modes of discovery

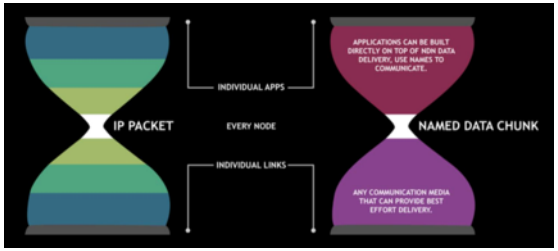


Agenda

Big Science Data



Global Science Collaborations



NDN for Science

Thanks to

Christos Papadopoulos,
Susmit Shannigrahi

imonga at es dot net

High-level objectives for scientific data: alignment with NDN approach

- Abstract the storage and network capability and location dependence from the **user-data interaction**
- Enable the ability for users to specify and retrieve **portions of data** the workflow needs
- Radically simplify how scientific users manage, move and manipulate large, distributed, science data repositories, but with **high-throughput end2end**
- Create a **secure, scalable framework** based on integrated data management and network transport

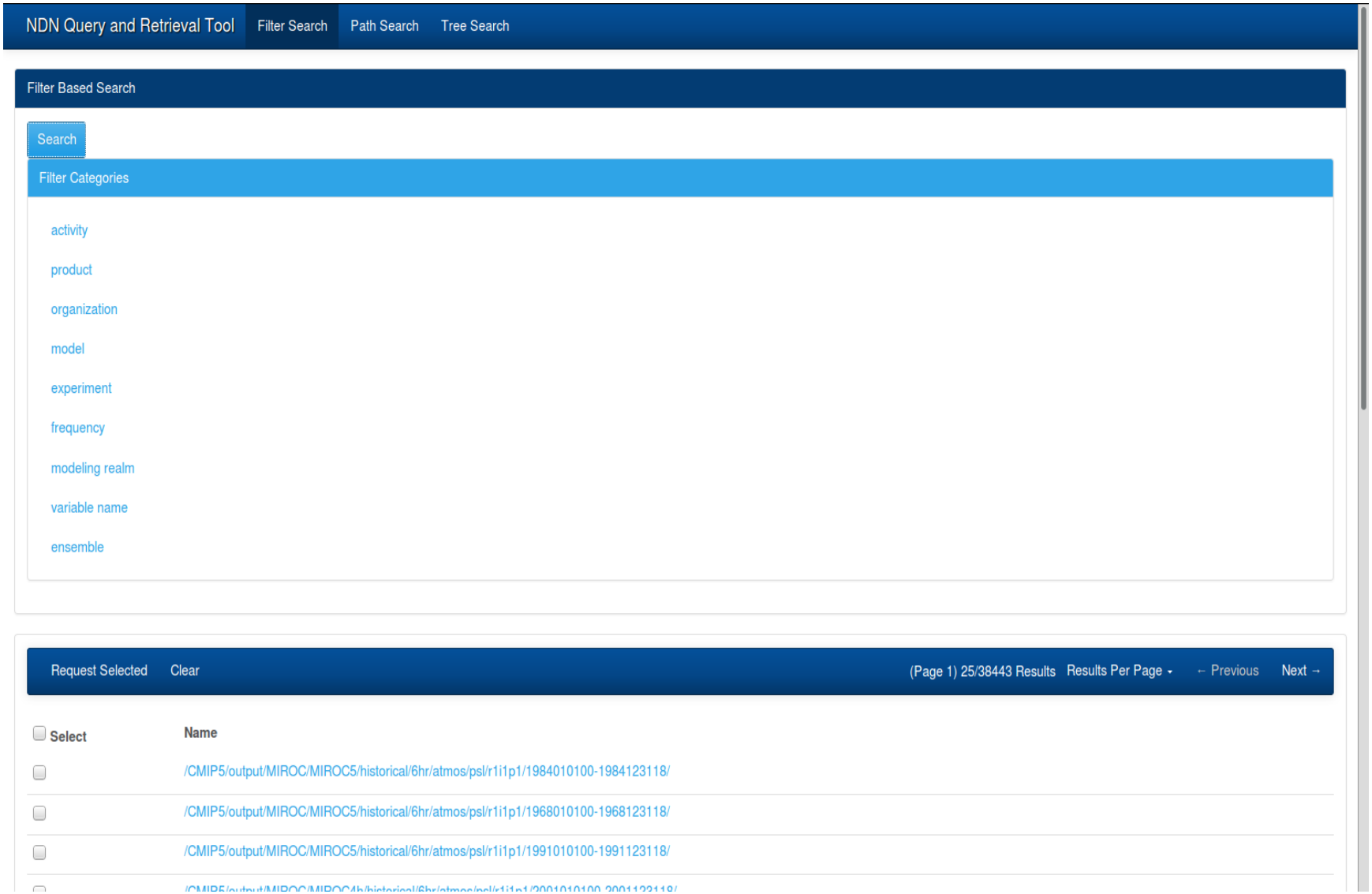
ESnet

imonga at es dot net

Challenge #1: Naming and Data Discovery

"Right off the bat, let's talk about name recognition."

Data Discovery UI



Data Discovery UI

- Three intuitive ways to search scientific data
  - Auto-complete, name component based search, and tree view
- Can work with any hierarchical datasets
  - We have two instances, for climate and HEP data
- Provides metadata browsing, subsetting, staging capabilities

ESnet

Challenge #2: Subsetting of data

ESnet

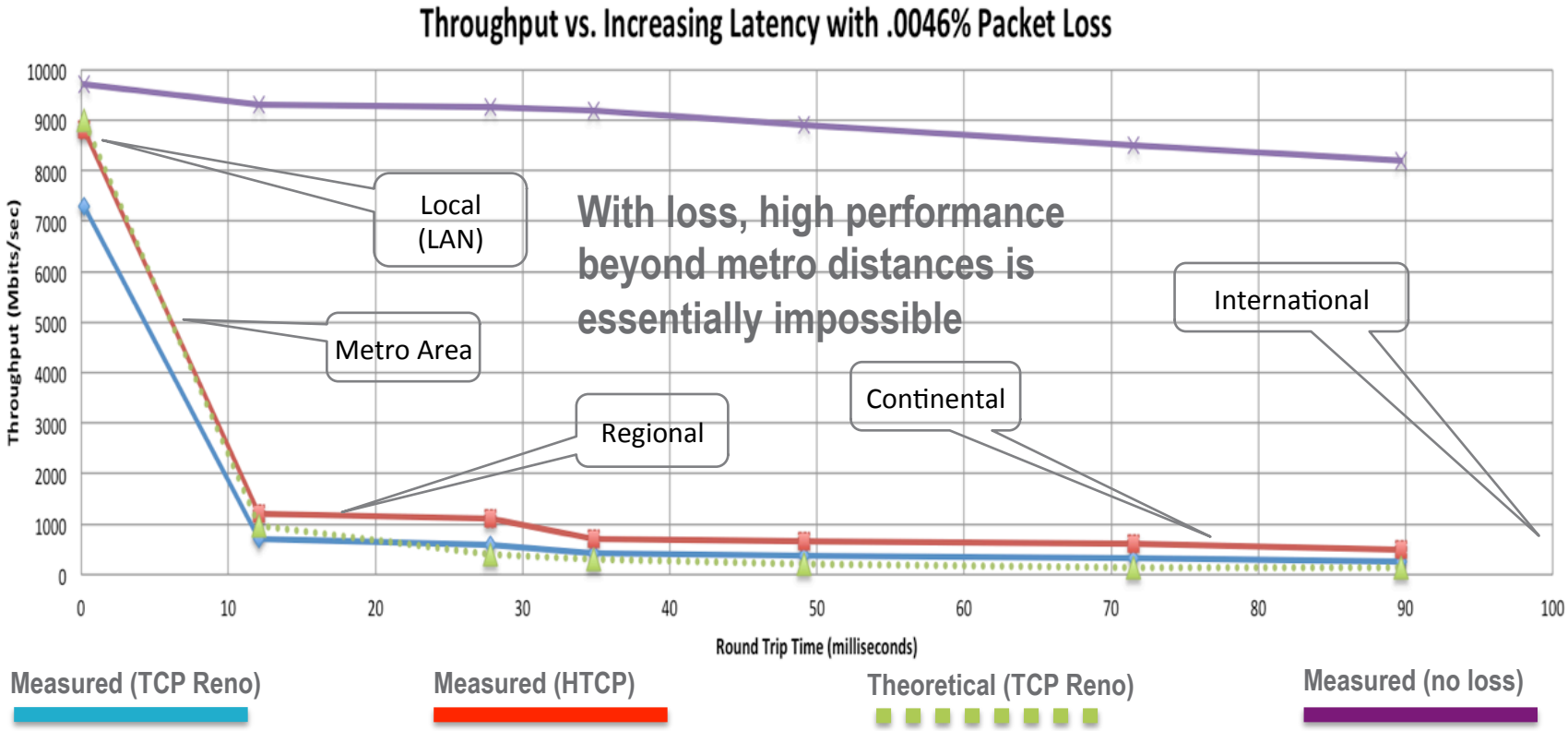
imonga at es dot net

Subsetting

- NDN names easily extend to support subsetting
- Add query parameters as a encoded name component
- Services can parse Interest name and perform intended action
  - ***Retrieval after subsetting*** is much economical than ***Subsetting after retrieval***

imonga at es dot net

Challenge #3: High performance end-2-end



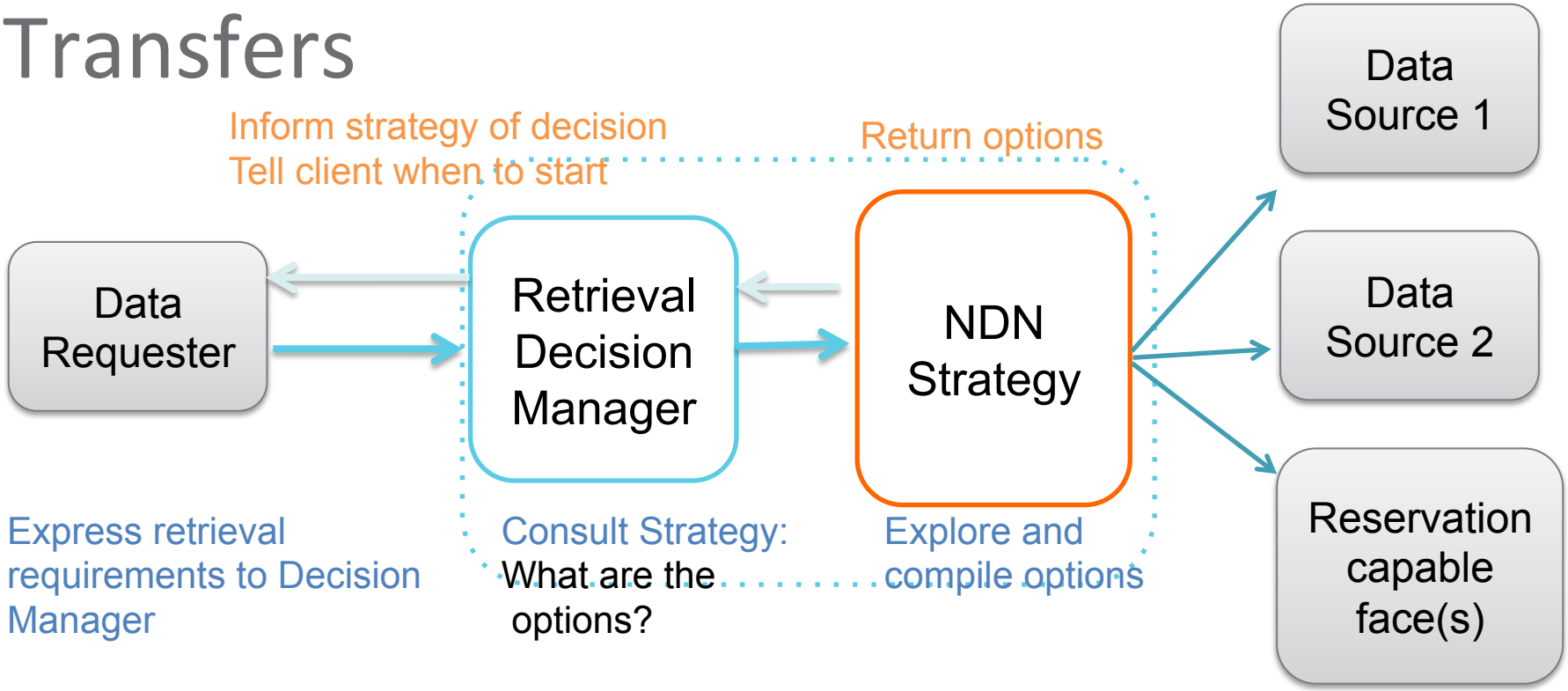
imonga at es dot net

NDN with OSCARS

- Some Data transfers require high bandwidth reserved paths
- We have integrated a NDN strategy with OSCARS
  - *A data retrieval Manager expresses special Interest to strategy layer*
  - *Strategy communicates with OSCARS to reserve a path*
  - *Interest/Data exchange uses the newly created path*
- Fully transparent to the application

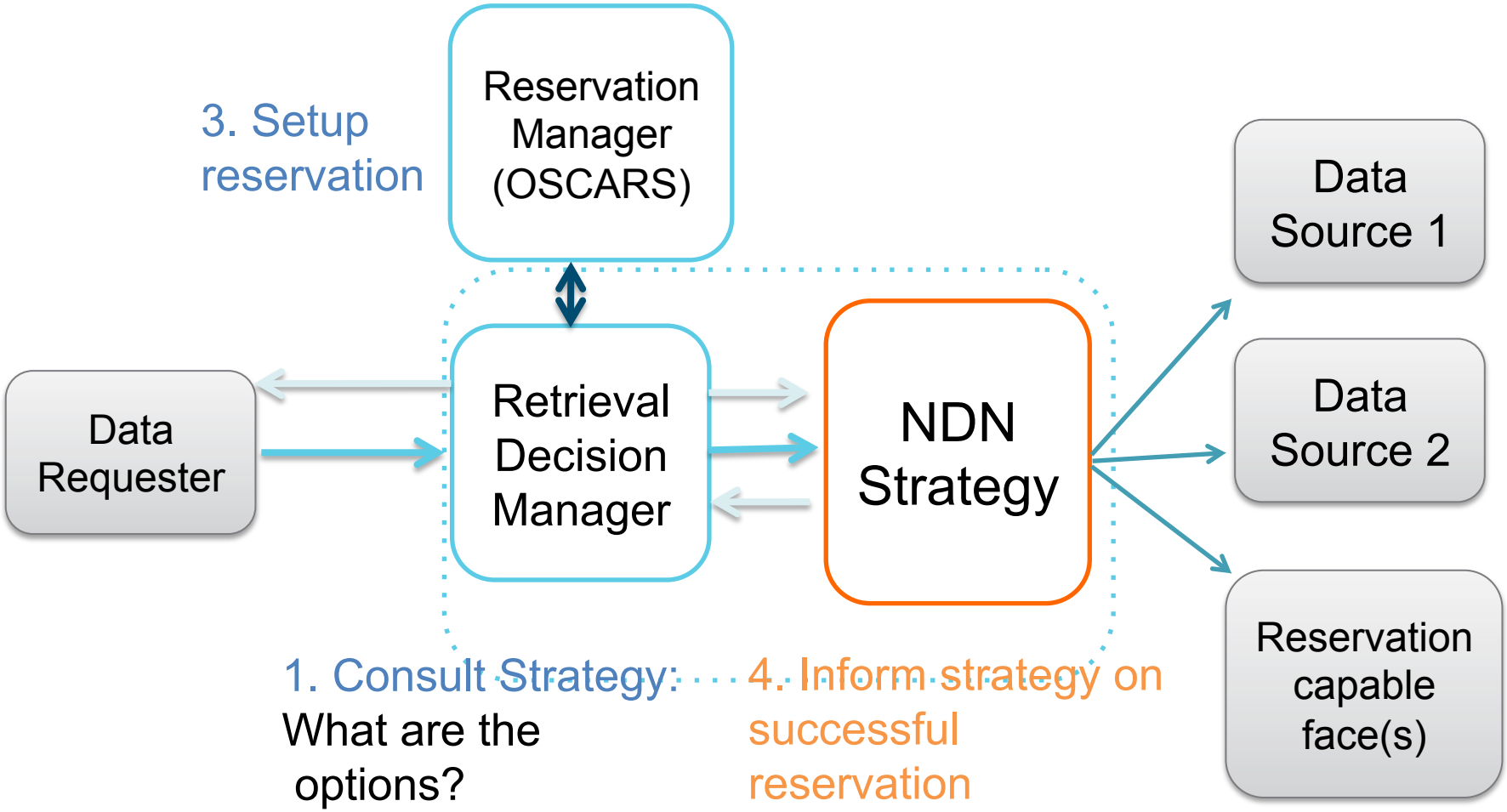
ESnet

NDN for Intelligent Data Transfers



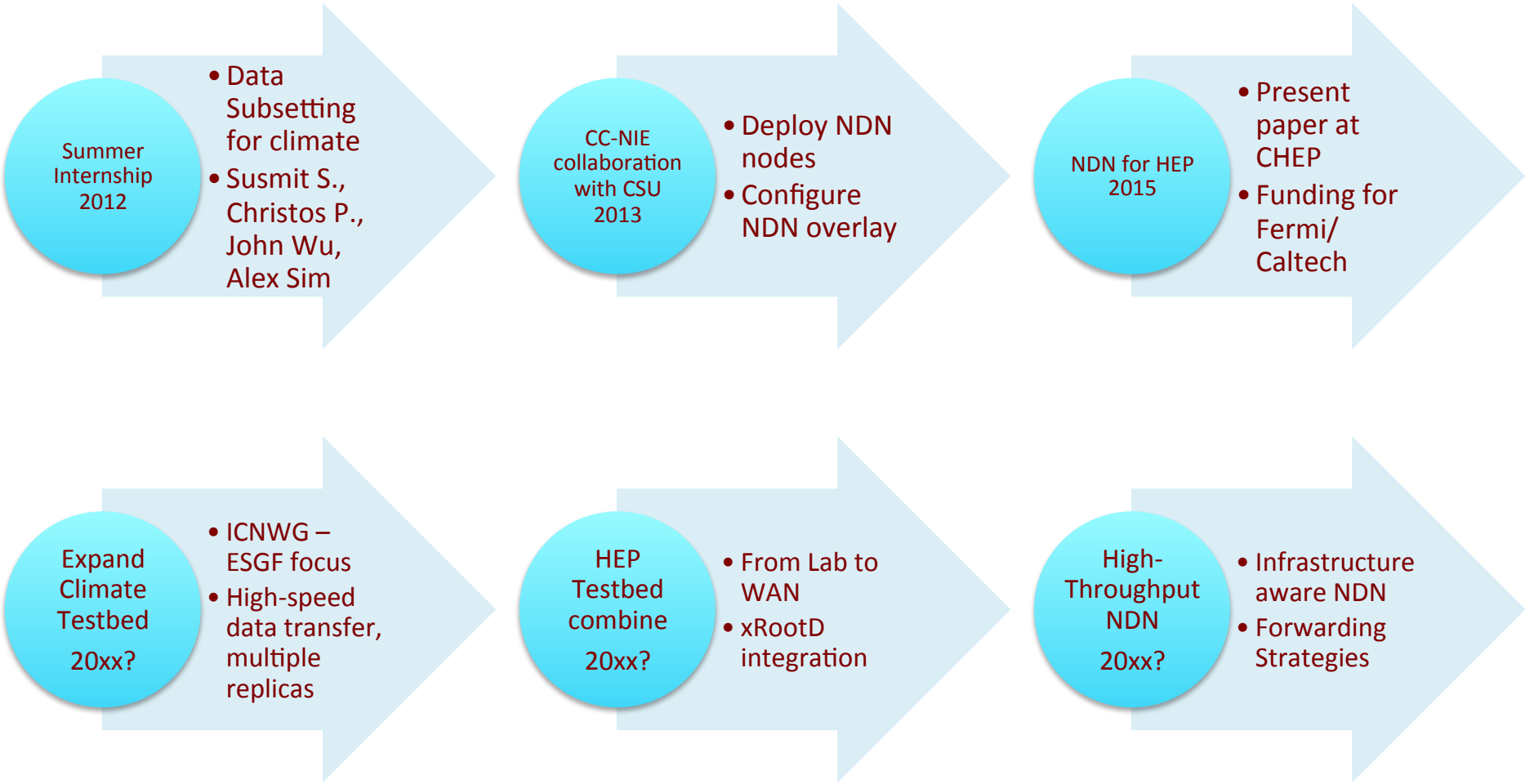
- Strategy for large scientific data transfers
- Retrieval Manager queries network for options
  - Makes a decision, informs strategy
  - Tells client to start retrieval

imonga at es dot net



imonga at es dot net

Roadmap for NDN Experimentation



ESnet

imonga at es dot net

Many unproven questions still…

- Where is the complexity being pushed to, and what needs to be done to manage that?
  - From the scientist to the network
- How can a network operator maintain, automate and operationally manage that complexity?
  - Think through the failure models
  - Think through performance models
- How does this work or compete with software scientists have already built to manage their data – what's the best way to integrate and/or migrate?

ESnet

The future is new data scientists!

17-year-old Brittany Wegner creates breast cancer detection tool that is 99% accurate on a minimally invasive, previously inaccurate test.

Machine Learning + Online Data + Cloud Computing

ESnet

Imonga at es dot net

Thank you!