



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



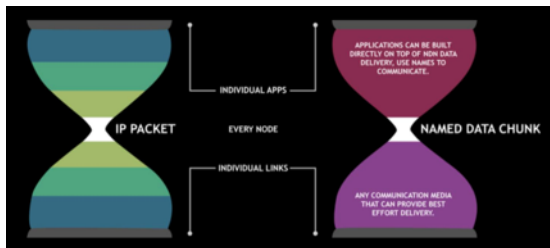
Agenda



Big Science Data



Global Science Collaborations



NDN for Science

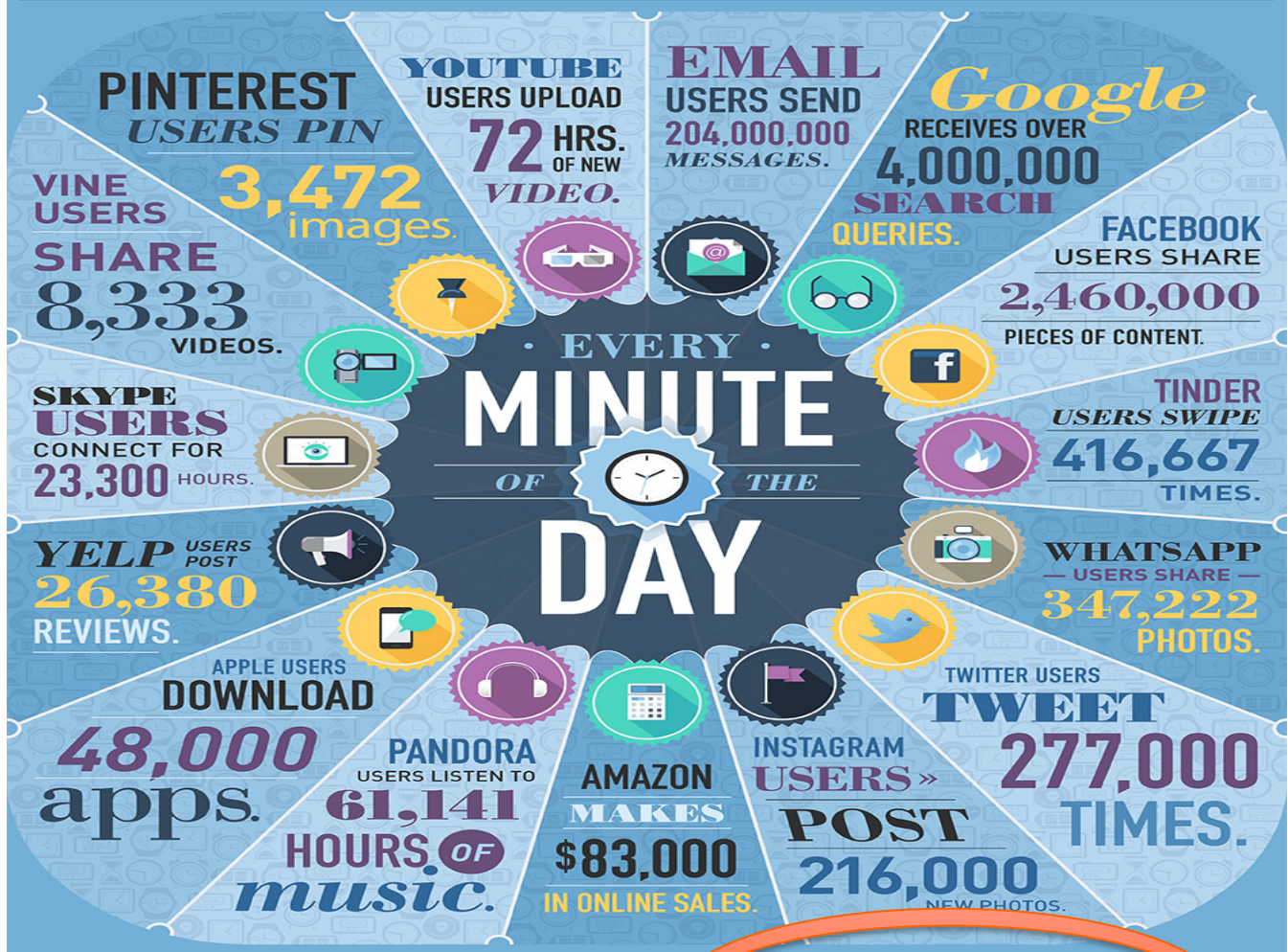
How big is data (visual comparison) *

- 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

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

Data is being created every minute of every day without us even noticing it. Given how much information is floating around these days, it's tempting to talk about big data only in terms of size. Big data describes the massive avalanche of digital activity pulsating through cables and airwaves, but it also describes all the things we were never able to measure before. With every status we share, every article we read or every photo we upload, we are creating a digital trail that tells a story. Below, we explore how much data is generated in one minute.



THE GLOBAL INTERNET POPULATION GREW **14.3%** FROM 2011 - 2013 AND NOW REPRESENTS

2.4 BILLION PEOPLE.

With each click, share and like, the world's data pool is expanding faster than we can comprehend. Businesses today are paying attention to scores of data sources to make crucial decisions about the future. The team at Domo can help your business make sense of this endless stream of data by providing executives with all their critical information in one intuitive platform. Domo delivers the insights you need to transform the way you run your business. Learn more at www.domo.com.

DOE Science “Apps”



Advanced Light Source

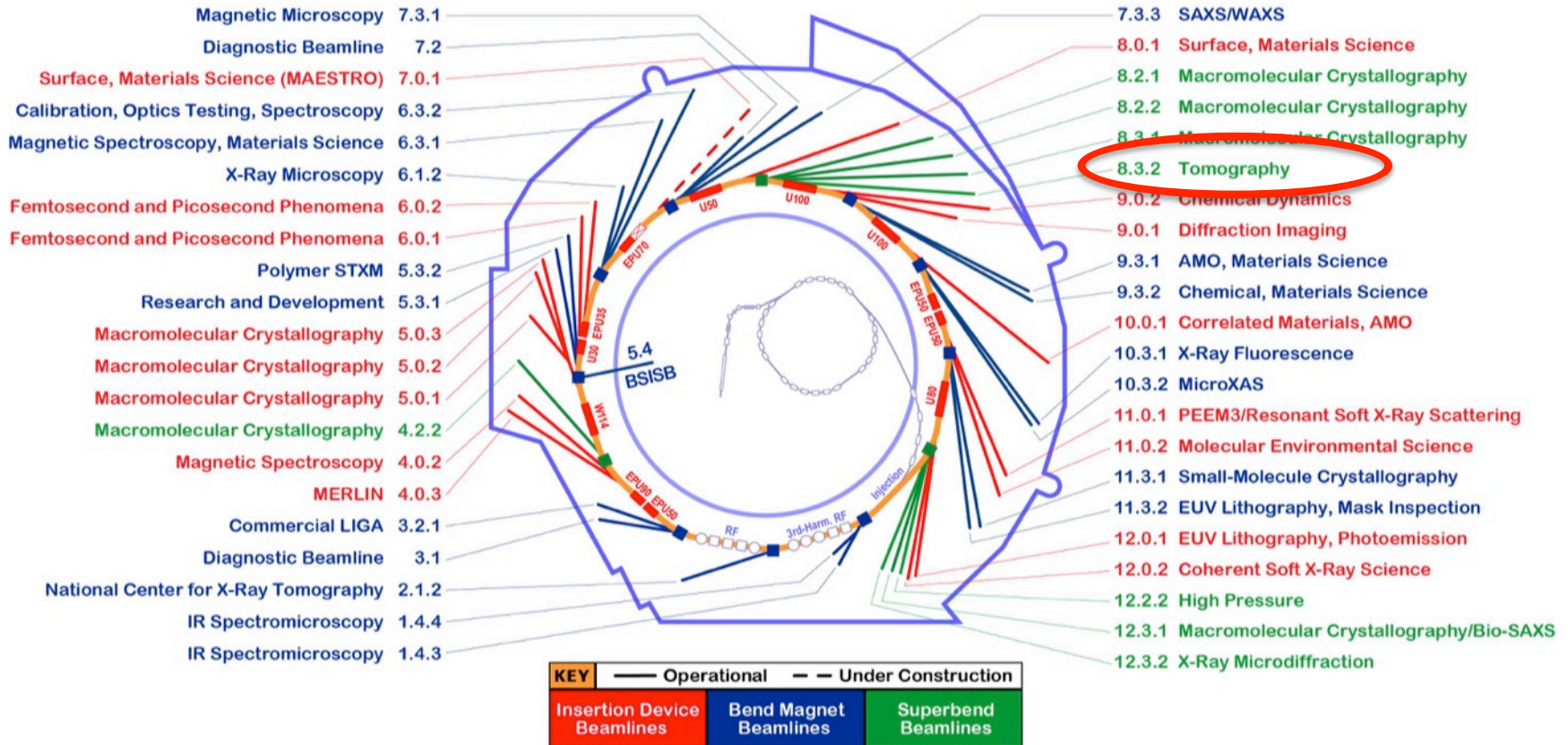


Advanced Light Source

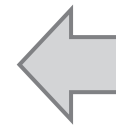
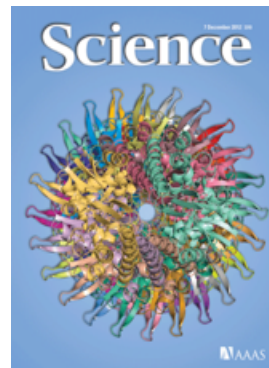
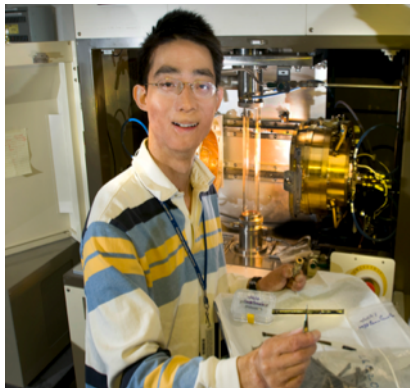
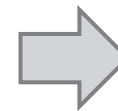
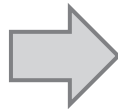
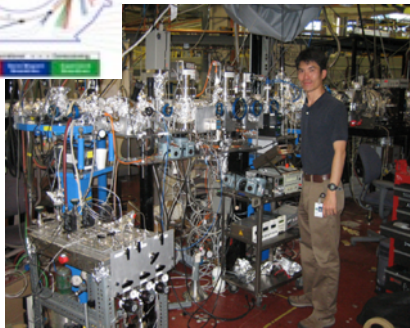
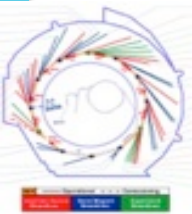


Advanced Light Source

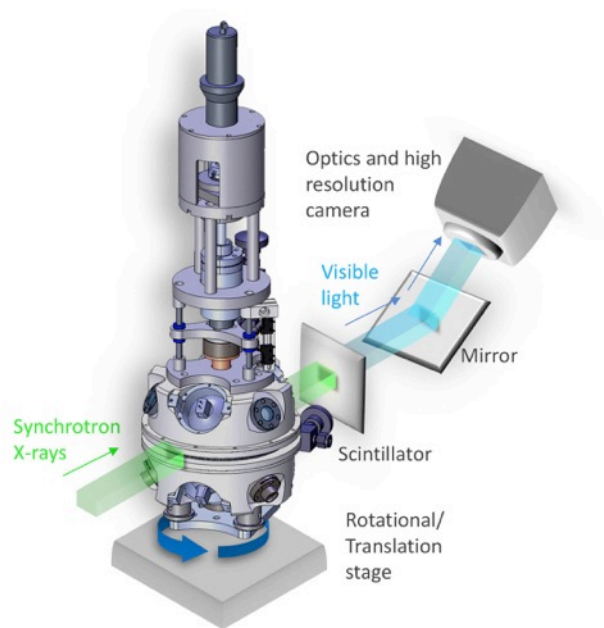
ALS Beamlines
January 2014



Scenario 1: All too common process of discovery

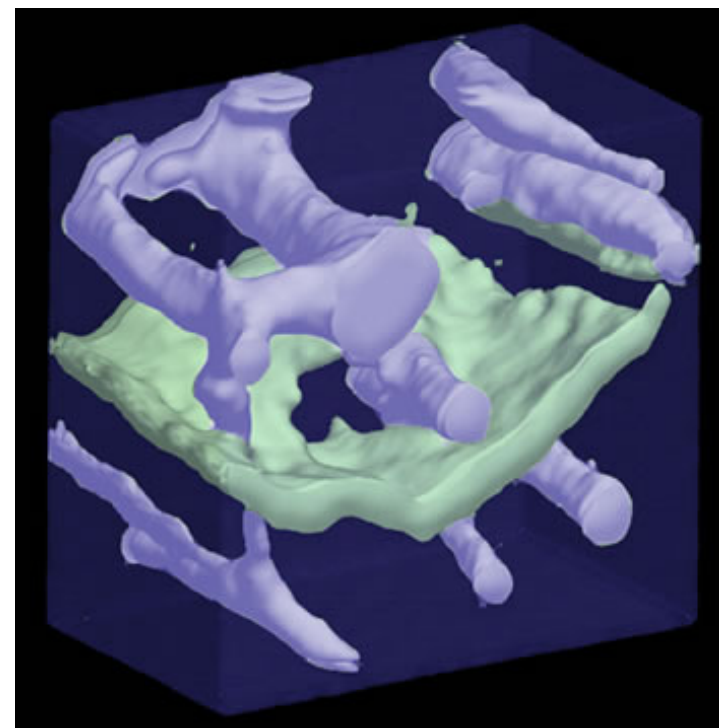


Beamline – Capture to Results



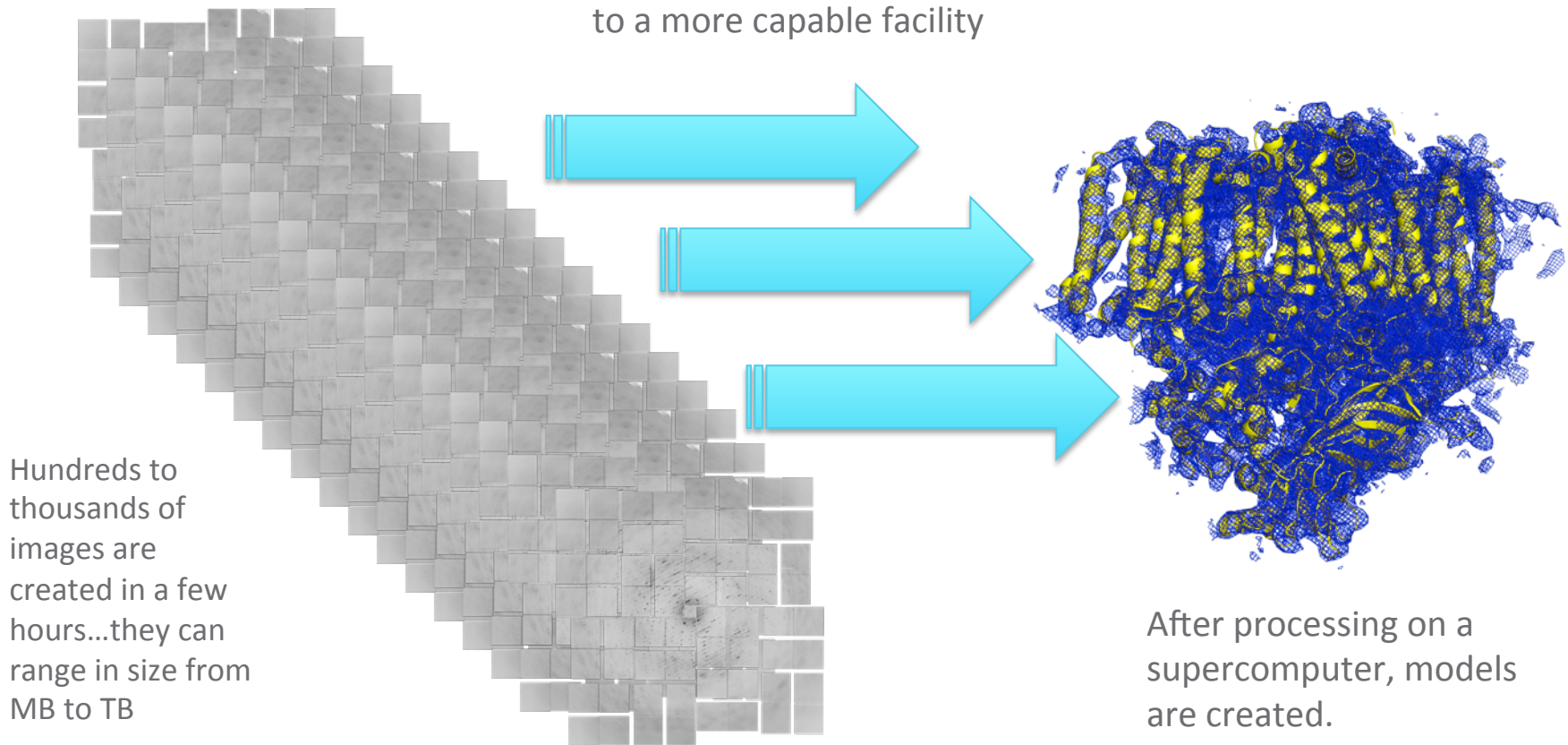
Basic Energy Sciences (BES) supports fundamental research to understand, predict, and ultimately control matter and energy at the electronic, atomic, and molecular levels in order to provide the foundations for new energy technologies and to support DOE missions in energy, environment, and national security.

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



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



Big Data vs. Big Data

Don't Forget:

Instagram Data produced/day
worldwide by millions of people

= **33 TB**

One Biology experiment at one
Beamline by a team of nine
scientists:

= **119 TB**

(Photosystem II X-Ray Study)

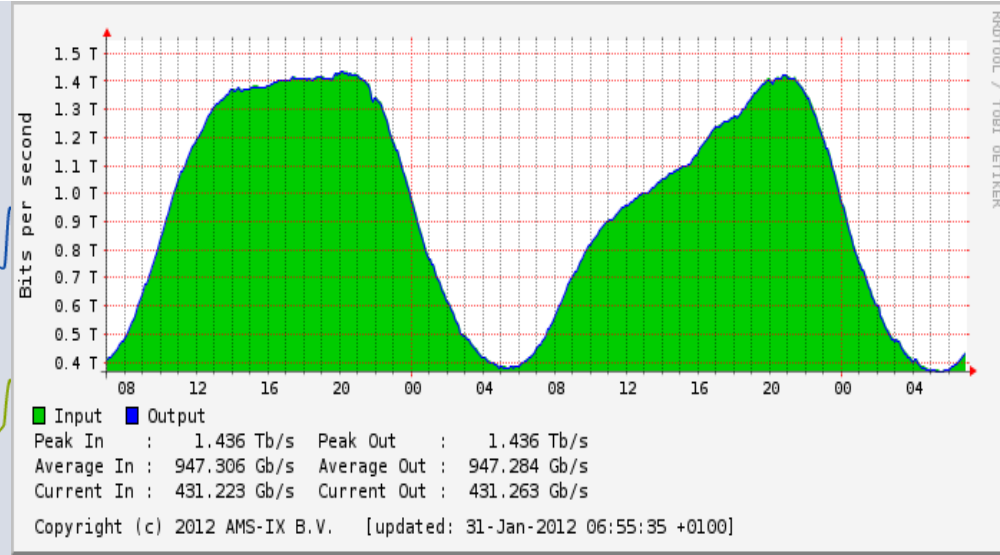
Big Science Data in Motion = Elephant Flow!

IoT watching LOL Cats = Mice flow!



Elephant Data vs. Mice Data Behavior

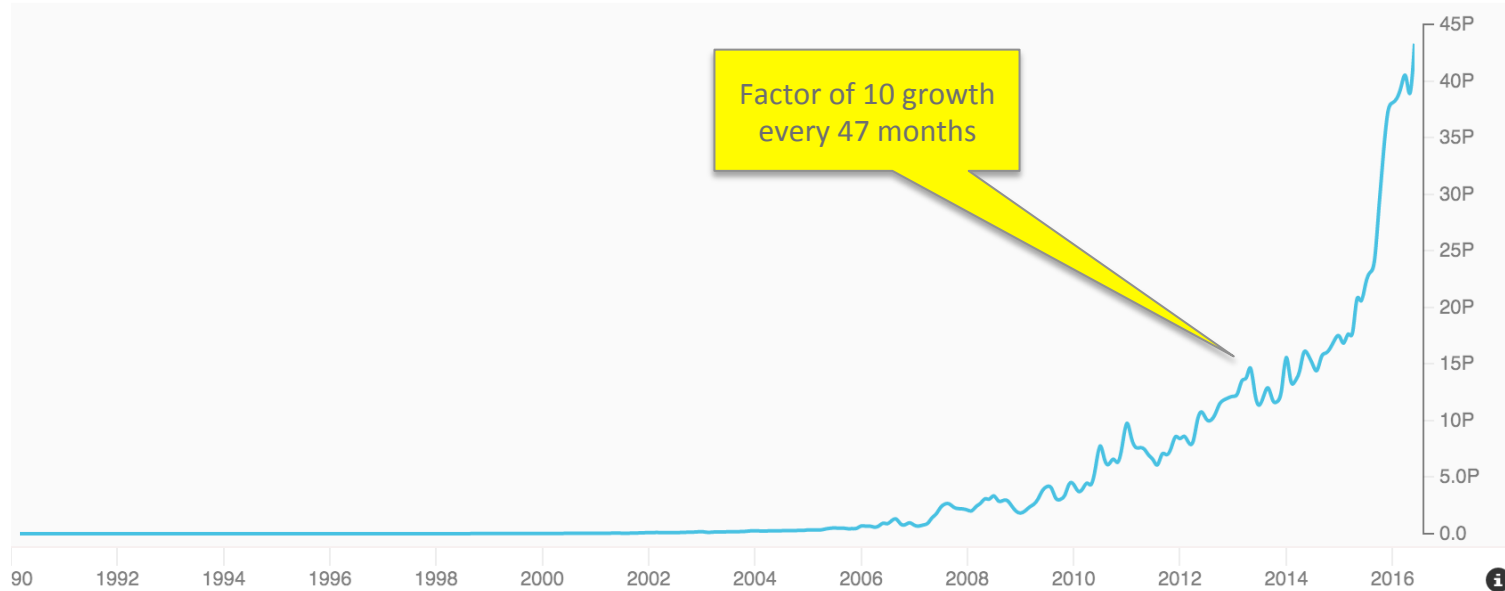
ESnet Traffic (last 24 hours)



Science Data Transferred Monthly by ESnet

Traffic Volume

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



◀ February 2016 ▶

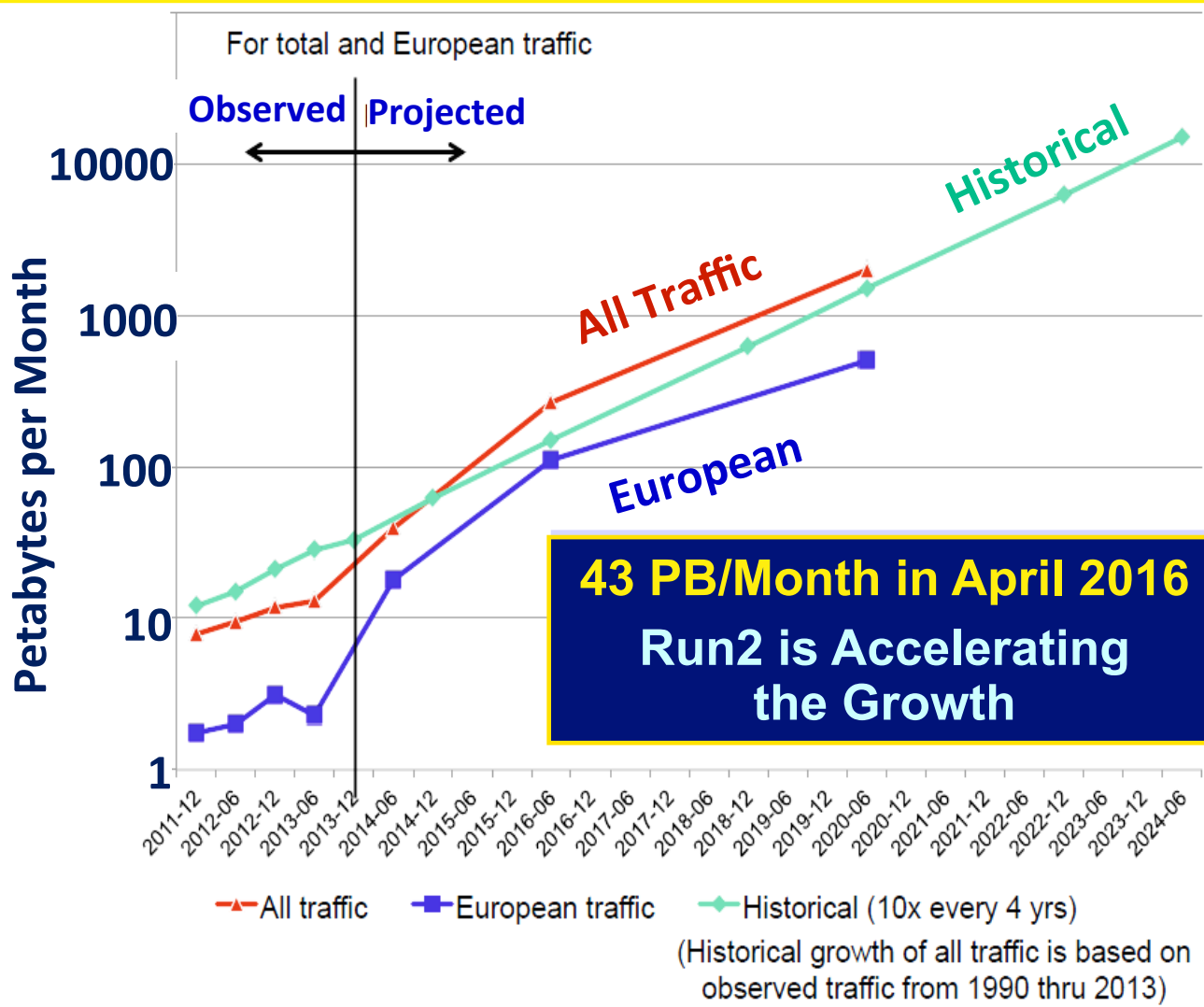
	Bytes	Percent of Total	One Month Change	One Year Change	
OSCARs	10.46 PB	25.2%	+0.0147%	+148%	Pt-to-pt circuits
LHCONE	11.22 PB	27.0%	+3.90%	+770%	LHCONE (T1-T1/2) traffic
Normal traffic	19.82 PB	47.8%	+9.49%	+77.7%	
Total	41.49 PB		+5.44%	+149%	

16 6/1/16

ESnet

imonga at es dot net

Traffic growth at blistering rates



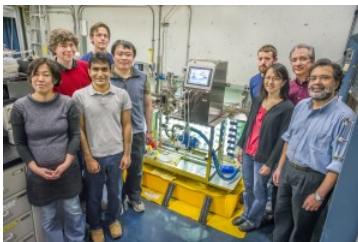
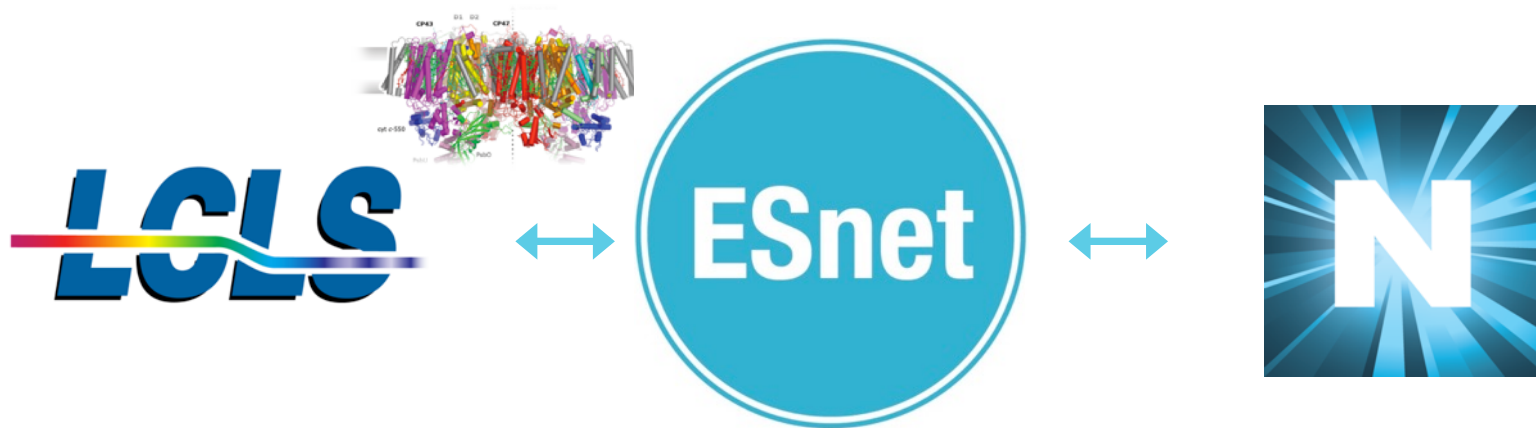
Projected Traffic Reaches
1 Exabyte Per Month. by ~2020
10 EB/Mo. by ~2024

Slide from
 Harvey Newman



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.



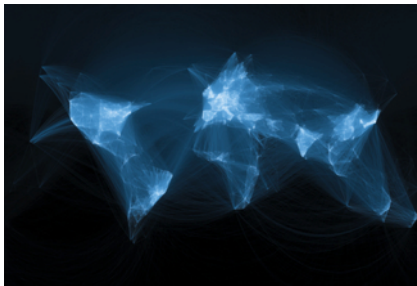
“Taking snapshots of photosynthetic water oxidation using femtosecond X-ray diffraction and spectroscopy,” *Nature Communications* 5, 4371 (9 July 2014)



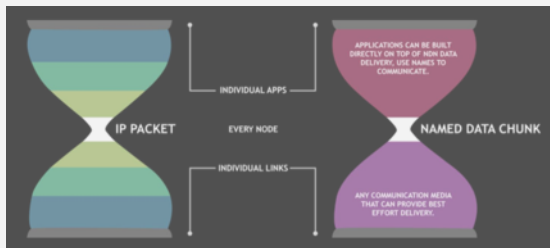
Agenda



Big Science Data



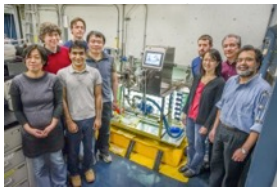
Global Science Collaborations



NDN for Science

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.

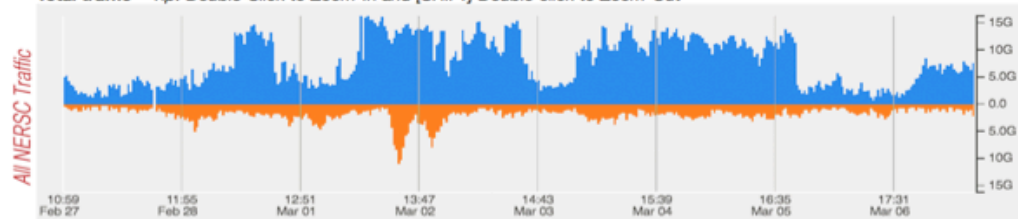


“Taking snapshots of photosynthetic water oxidation using femtosecond X-ray diffraction and spectroscopy,”
Nature Communications 5, 4371 (9 July 2014)

From : Wed Feb 27 10:59:00 2013 To : Thu Mar 7 10:59:00 2013

■ To site ■ From site

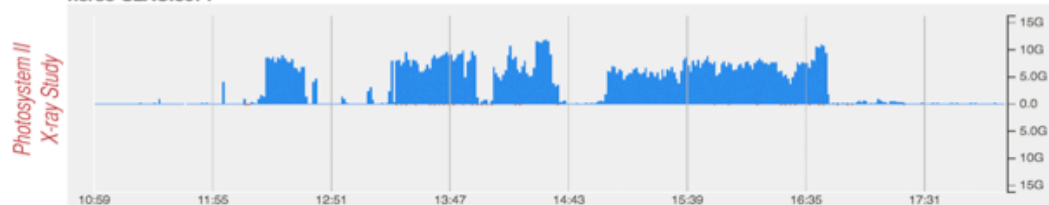
Total traffic Tip: Double Click to Zoom-In and [SHIFT] Double click to Zoom-Out



Traffic split by : 'Autonomous System (origin)'

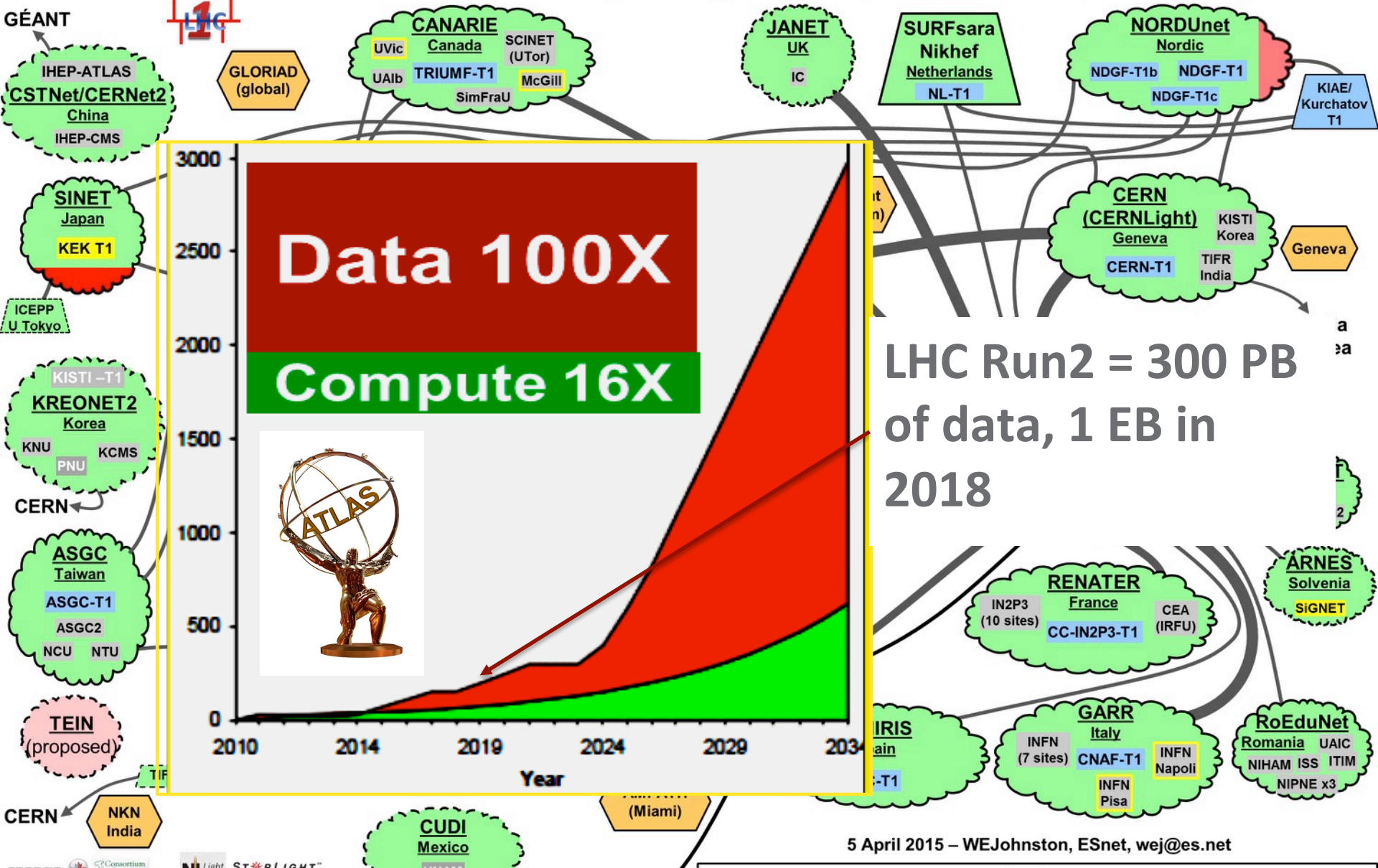
50TB moved a night

nersc-SLAC:3671



ESnet

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



5 April 2015 – WEJohnston, ESnet, wej@es.net

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

Use Case 3: Worldwide Earth System Grid Federation Sites



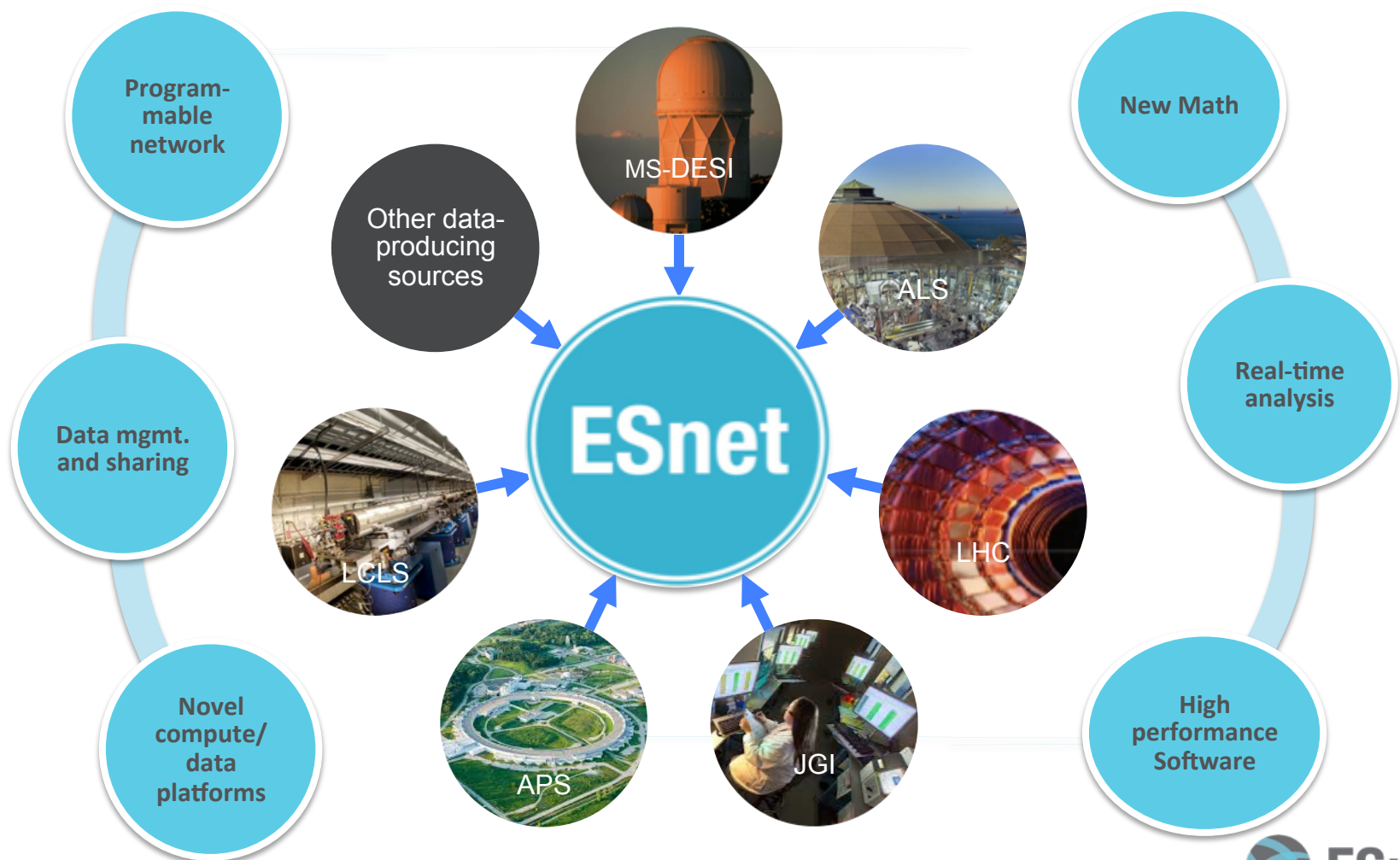
Use Case Galore

#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.

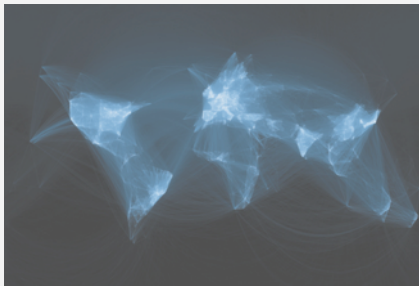
Superfacility Vision: A **network** of connected facilities, software and expertise to enable new modes of discovery



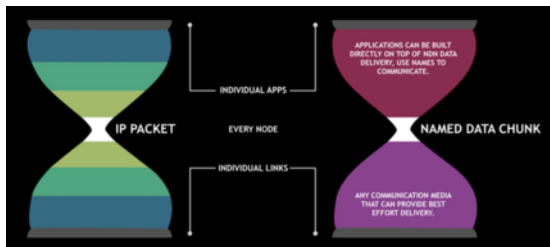
Agenda



Big Science Data



Global Science Collaborations



NDN for Science

Thanks to

Christos Papadopoulos,
Susmit Shannigrahi



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

Challenge #1: Naming and Data Discovery

© MARK ANDERSON

WWW.ANDERTOONS.COM



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

Data Discovery UI

NDN Query and Retrieval Tool

Filter Search

Path Search

Tree Search

Filter Based Search

Search

Filter Categories

activity
product
organization
model
experiment
frequency
modeling realm
variable name
ensemble

Request Selected Clear

(Page 1) 25/38443 Results Results Per Page - Previous Next ->

<input type="checkbox"/> Select	Name
<input type="checkbox"/>	/CMIP5/output/MIROC/MIROC5/historical/6hr/atmos/psl/r1i1p1/1984010100-1984123118/
<input type="checkbox"/>	/CMIP5/output/MIROC/MIROC5/historical/6hr/atmos/psl/r1i1p1/1968010100-1968123118/
<input type="checkbox"/>	/CMIP5/output/MIROC/MIROC5/historical/6hr/atmos/psl/r1i1p1/1991010100-1991123118/
<input type="checkbox"/>	/CMIP5/output/MIROC/MIROC4h/historical/6hr/atmos/psl/r1i1p1/2001010100-2001123118/



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

Challenge #2: Subsetting of data

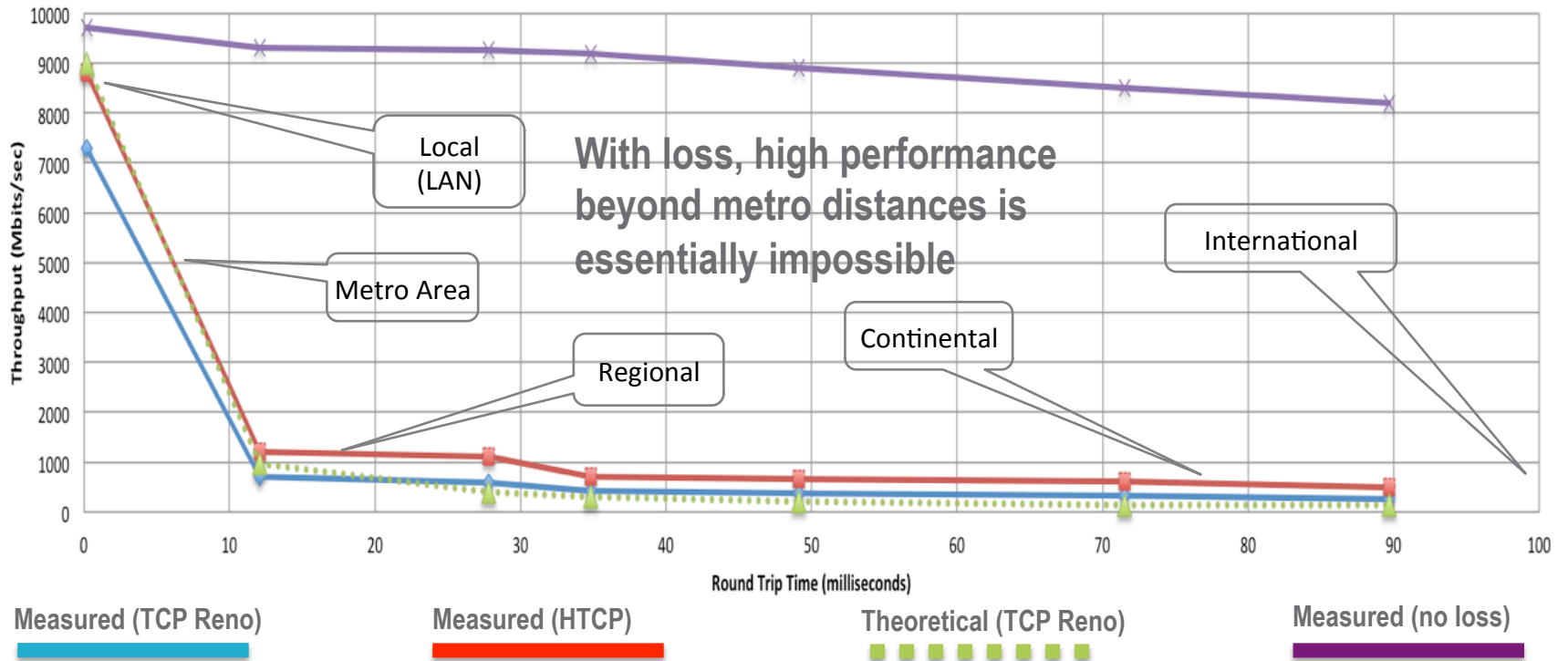


Subsetting

- 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*

Challenge #3: High performance end-2-end

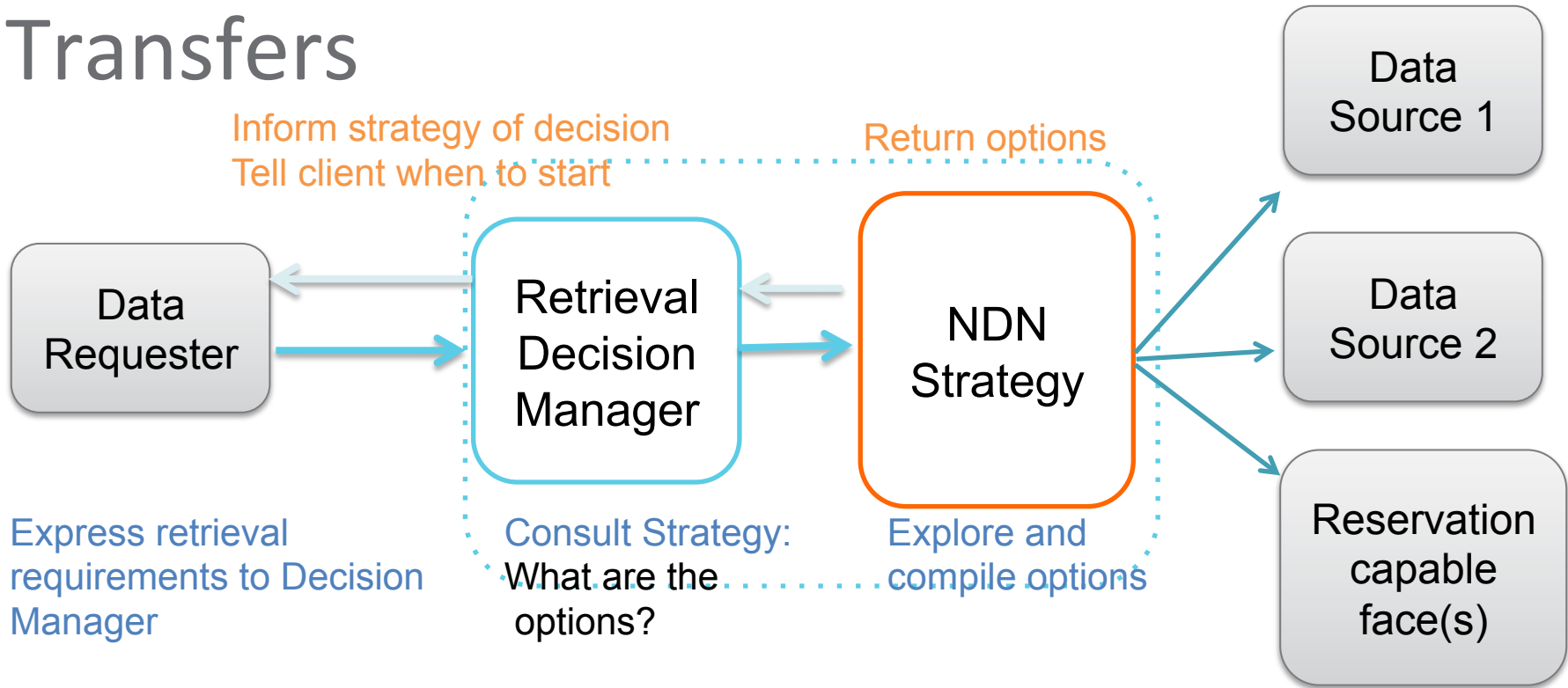
Throughput vs. Increasing Latency with .0046% Packet Loss



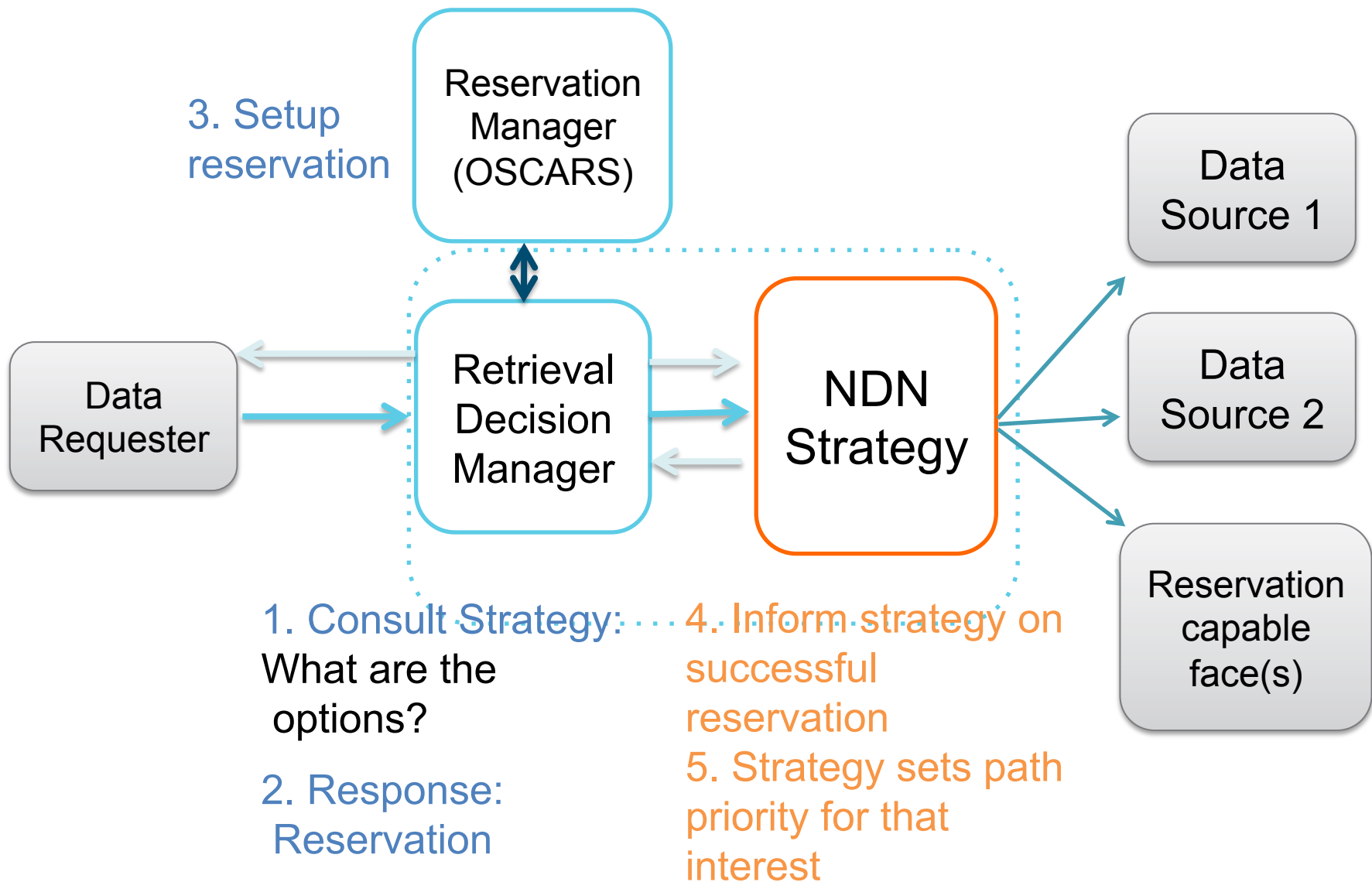
NDN with OSCARS

- 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

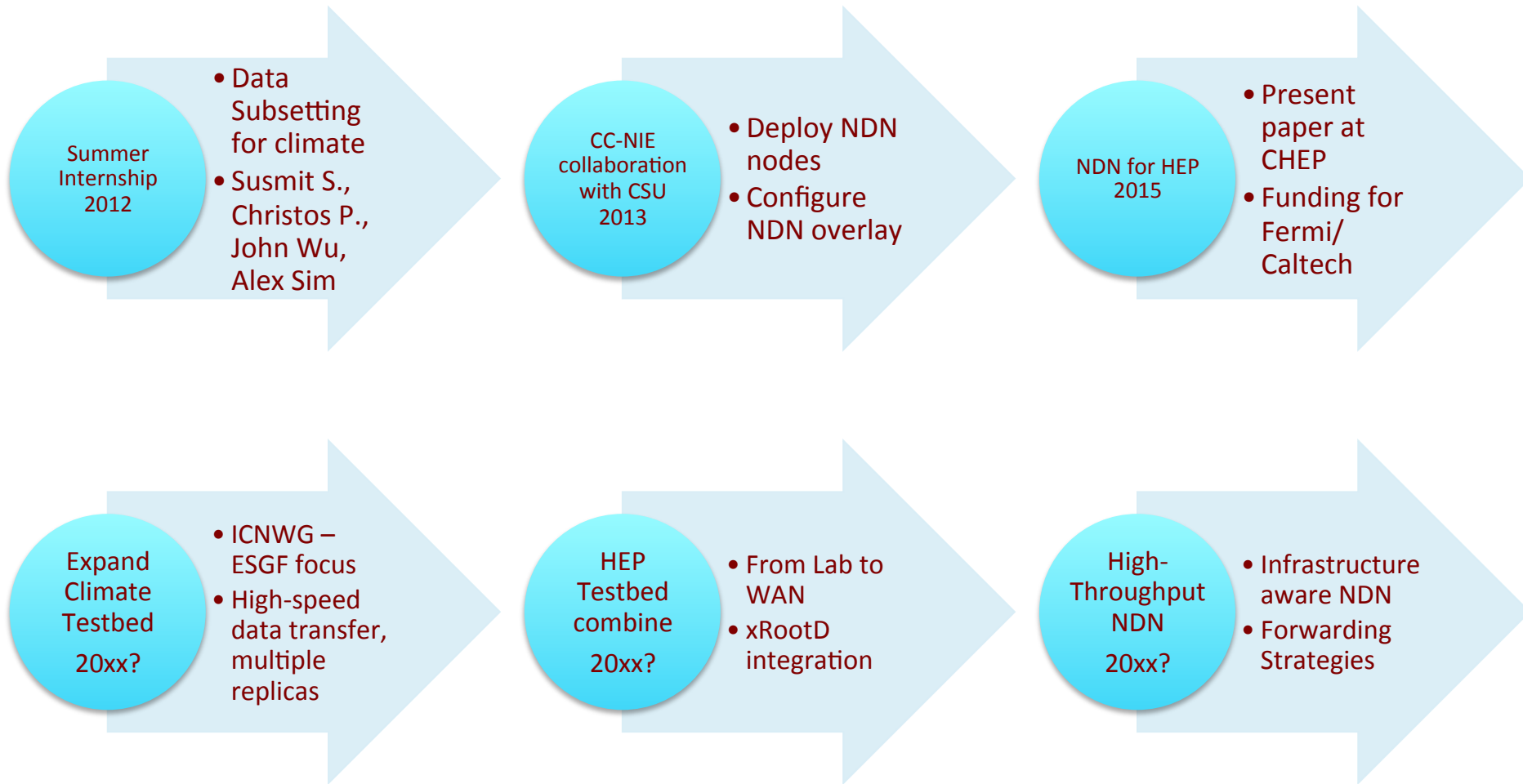
NDN for Intelligent Data Transfers



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



Roadmap for NDN Experimentation



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?

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



Imonga at es dot net

Thank you!