Published Online on 12 June 2018

Proc Indian Natn Sci Acad **84** No. 2 June 2018 pp. 359-370

DOI: 10.16943/ptinsa/2018/49338

Review Article

Big-Data Science: Infrastructure Impact

INDER MONGA and PRABHAT
Berkeley and NERSC

Introduction

The nature of science is changing dramatically, from single researcher at a lab or university laboratory working with graduate students to a distributed multi-researcher consortiums, across universities and research labs, tackling large scientific problems. In addition, experimentalists and theorists are collaborating with each other by designing experiments to prove the proposed theories. 'Big Data' being produced by these large experiments have to verified against simulations run on High Performance Computing (HPC) resources.

The trends above are pointing towards

a. Geographically dispersed experiments (and associated communities) that require data being moved across multiple sites. Appropriate mechanisms and tools need to be employed to move, store and archive datasets from such experiments.

b. Convergence of simulation (requiring High Performance Computing) and Big Data Analytics (requiring advanced on-site data management techniques) into a small number of High Performance Computing centers. Such centers are key for consolidating software and hardware infrastructure efforts, and achieving broad impact across numerous scientific domains.

The trends indicate that for modern science and scientific discovery, infrastructure support for handling both large scientific data as well as high-performance computing is extremely important. In addition, given the distributed nature of research and big-team science, it is important to build infrastructure, both hardware and software, that enables sharing across institutions, researchers, students, industry and academia. This is the only way that a nation can maximize the research capabilities of its citizens while maximizing the use of its investments in computer, storage, network and experimental infrastructure.

This chapter introduces infrastructure requirements of High-Performance Computing and Networking with examples drawn from NERSC and ESnet, two large Department of Energy facilities at Lawrence Berkeley National Laboratory, CA, USA, that exemplify some of the qualities needed for future Research & Education infrastructure.

Most scalable Deep-learning Implementation

National Energy Research Scientific Computing Center (NERSC) reported in their communication dated 28 August 2017, that a collaborative effort between Intel, NERSC and Stanford has delivered the first 15-petaflops deep learning software running on HPC platforms and is, according to the authors of the paper (and to the best of their knowledge), currently the most scalable deep-learning implementation in the world. The work described in the paper, Deep Learning at 15PF: Supervised and Semi-Supervised Classification for Scientific Data (https://arxiv.org/abs/1708.05256), reported that a Cray XC40 system with a configuration of 9,600 self-hosted 1.4GHz Intel Xeon Phi Processor 7250 based nodes achieved a peak rate between 11.73 and 15.07 petaflops (single-precision) and an average sustained performance of 11.41 to 13.47 petaflops when training on physics and

*Author for Correspondence: E-mail: imonga@es.net

climate based data sets using Lawrence Berkeley National Laboratory's (Berkeley Lab) NERSC (National Energy Research Scientific Computing Center) Cori Phase-II supercomputer. The group utilized an amalgamation of Intel Caffe, Intel Math Kernel Library (Intel MKL), and Intel Machine Learning Scaling Library (https://github.com/01org/MLSL) (Intel MLSL) software to achieve this scalability and performance.

High-Performance Computing

As one of the world's premier supercomputing centers, NERSC supports perhaps the largest and most diverse research community of any high-performance computing facility, providing large-scale, state-of-the-art computing for DOE's unclassified research programs. More than 6,000 scientists worldwide use NERSC to conduct basic and applied research in energy production and conservation, climate change, environmental science, materials research, chemistry, fusion energy, astrophysics and other areas related to the mission of the DOE Office of Science. Fig. 3.1, provides a brief overview of the hardware resources at NERSC. Two major supercomputing platforms are operational at any point in time; currently we host two petaflop class systems: Cray XE6 system (Hopper) and Cray XC30 (Edison). These tightly coupled systems feature a high performance interconnect, and a fast, distributed parallel filesystem. Relatively higher capacity, but lower bandwidth project and archival systems are available to users for longer term retention of data. NERSC has experimented with installing dedicated, smaller-scale clusters for handling data-intensive workloads of specific domain science communities. Finally, in order to move data efficiently between supercomputing centers, dedicated data transfer nodes provide an end-point for both 10G and 100G ESnet connections.

In 2016, NERSC will install Cori, a Cray XC40 system. This system will provide unified resources for handling both HPC, as well as data-centric workloads. In the HPC space, NERSC is making a major push to port applications to energy-efficient, many-core architectures with the NESAP (NERSC Exascale Science Application Program); teaming up talented post-docs with strategic applications. In the Big Data space, NERSC is innovating on a number of fronts: we are utilizing the Datawarp technology to

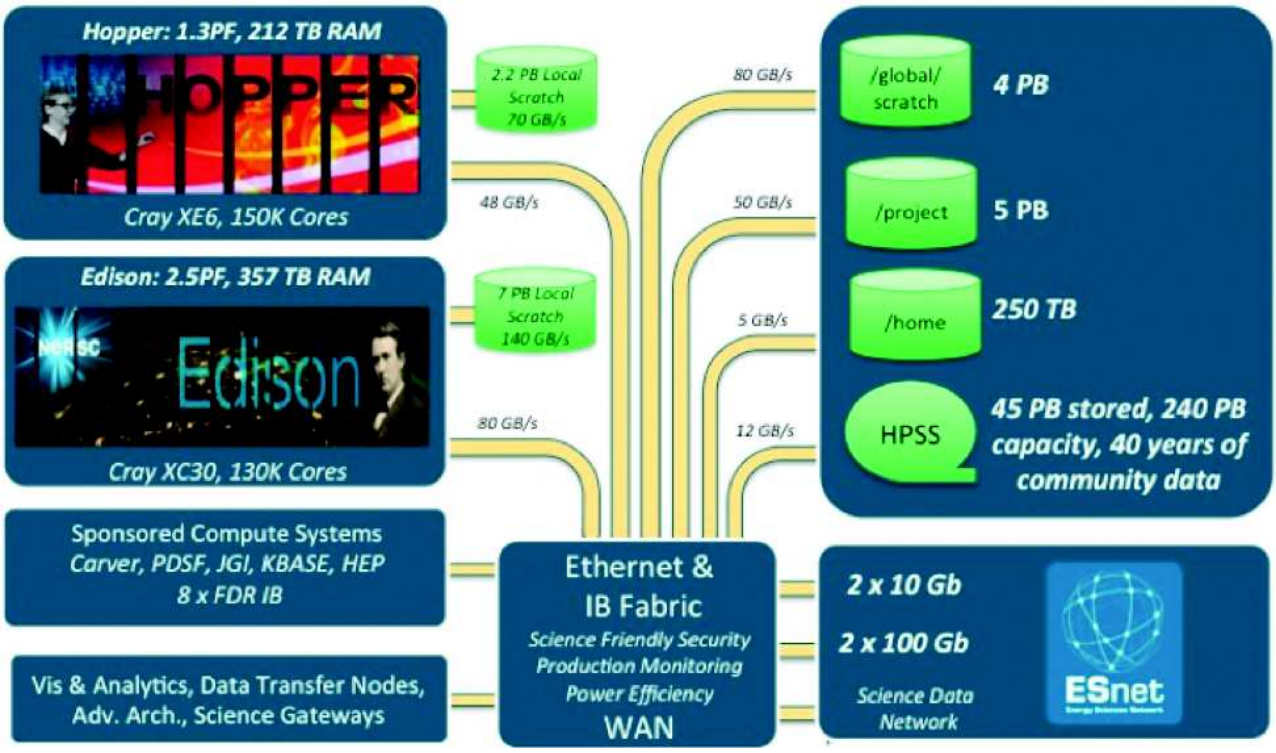

Fig. 3.1: Overview of NERSC systems (circa October 2015)

provide users with access to extremely high bandwidth and low-latency NVRAM storage; we are configuring our batch system to provide real-time, interactive, serial and high-throughput queues; we are enabling compute nodes on the system to have external connectivity, and we are enabling custom user-environments through Docker-like containers.

In terms of HPC software, NERSC provides a broad portfolio of compilers, code development tools, domain-specific application codes, programming libraries, performance and debugging tools. In the Big Data space, NERSC provides software capabilities in the areas of Data analytics, Data Management, Work-flows, Data Transfer, Data Access and Visualization. Documentation on all of these capabilities are provided at www.nersc.gov and regular outreach and training events are conducted to keep the scientific community abreast of latest technologies.

NERSC gathers HPC, data and services requirements from the science community in many ways. Chief among them are the program requirements reviews held with each of the six offices within the DOE Office of Science. This ongoing series of reviews brings together DOE program managers, leading domain scientists and NERSC staff to derive each scientific community's future HPC needs. The results of the reviews include requirements for computing, storage and services five years out. Each review report also contains a number of significant observations, topics of keen interest to the review participants. These results help DOE and NERSC plan for future systems and HPC service offerings.

NERSC is much more than just a collection of computers, servers, routers and software tools. One of its most valuable attributes is its staff, a talented group of computer scientists, mathematicians, engineers and support personnel. More than 50 percent of NERSC staff hold advanced degrees in a scientific or technical field. And collaboration-aka "team science," a concept pioneered by Berkeley Lab founder Ernest O. Lawrence in 1931-is a cornerstone of NERSC's philosophy, both internally and through its engagements with the broader science community.

Impact of Computing on Science

Theory, Experiment, Simulation and Data-Driven Discovery are now widely accepted as the four paradigms of modern science. Simulation and High Performance Computing go hand-in-hand; all natural or man-made systems require higher fidelity, either in terms of the spatial/temporal resolutions, or in terms of the physical processes being modeled. HPC has had a broad impact across a number of domains, as highlighted by the following brief examples:

- NOAA routinely use HPC resources for making regional weather forecasts over the US and UK respectively
- Major aircraft manufacturers (Boeing, Airbus) use HPC to create, and simulate digital models of planes before fabrication
- NASA utilizes HPC to explore space shuttle and spacecraft design for both robotic and manned missions
- DOE utilizes HPC to simulate next generation Tokamak reactors for exploring the promise of fusion energy
- NSF utilizes HPC to conduct simulations of earthquakes along various fault lines in California, and impact on local economies
- Several firms on the the Wall Street utilize HPC resources to enable high frequency trading
- Intelligence agencies utilize HPC resources to find patterns and anomalies in unstructured data.

Big Data has its origins in the commercial world. Internet-driven companies such as Google, Facebook and Twitter need to be able to analyze massive amounts of user data, and find mechanism to add value to their user's online experience, as well as monetize user behavior to generate a revenue stream. Major investments have been made by these firms in data-centers throughout the globe, and an associated software stack.

In the remainder of this article, we will focus on success stories from NERSC, that highlight the kind of progress that can be achieved in basic sciences through investments in HPC and Big Data resources.

Scientific Success Stories

Over the past 40 years of NERSC's history, we have

witnessed the evolution of High Performance Computing from Terascale to Petascale and now en route to Exascale class systems. HPC has been successfully applied to simulate the evolution of the universe, model supernova explosions, model climate change, simulate carbon sequestration, perform quantum mechanical simulations of various materials and simulate experiments such as the Large Hadron Collider on its search for sub-atomic constituents.

Big Data Analytics is a relatively recent trend at NERSC, having gained prominence over the last 5 years. Major projects include high throughput pipelines for genome assembly, automated candidate identification in astronomy images, 3D reconstruction of light source data, interactive exploration of high energy physics experiments and so on. We also observe the trend of the integration of observational data with simulations: a classic example is the production of climate 'reanalysis' datasets, which use a climate model to interpolate satellite and weather station datasets.

Annually, NERSC users produce over 1900 publications in top-tier scientific venues such as *Nature, Science, PNAS,* etc. NERSC also has a rich history of contributions to a number of Nobel Prizes:

- 2015 Nobel Prize in Physics on discovery of neutrino oscillations
- 2013 Nobel Prize in Chemistry on development of multi-scale models for complex chemical systems
- 2011 Nobel Prize in Physics on measuring the acceleration of cosmic expansion
- 2007 Nobel Peace Prize on characterization of climate change
- 2006 Nobel Prize in Physics on Cosmic Microwave Background Radiation

In the next two subsections, we briefly comment on science stories which show the successful application of HPC, as well as Big Data Analytics methods to further scientific discovery.

Characterizing Extreme Weather in a Changing Climate

Not long ago, it would have taken several years to run a high-resolution simulation on a global climate model. But using supercomputing resources at NERSC, in 2014 Berkeley Lab climate scientist Michael Wehner was able to complete a run in just three months. What he found was that not only were the simulations much closer to actual observations, but the high-resolution models were far better at reproducing intense storms, such as hurricanes and

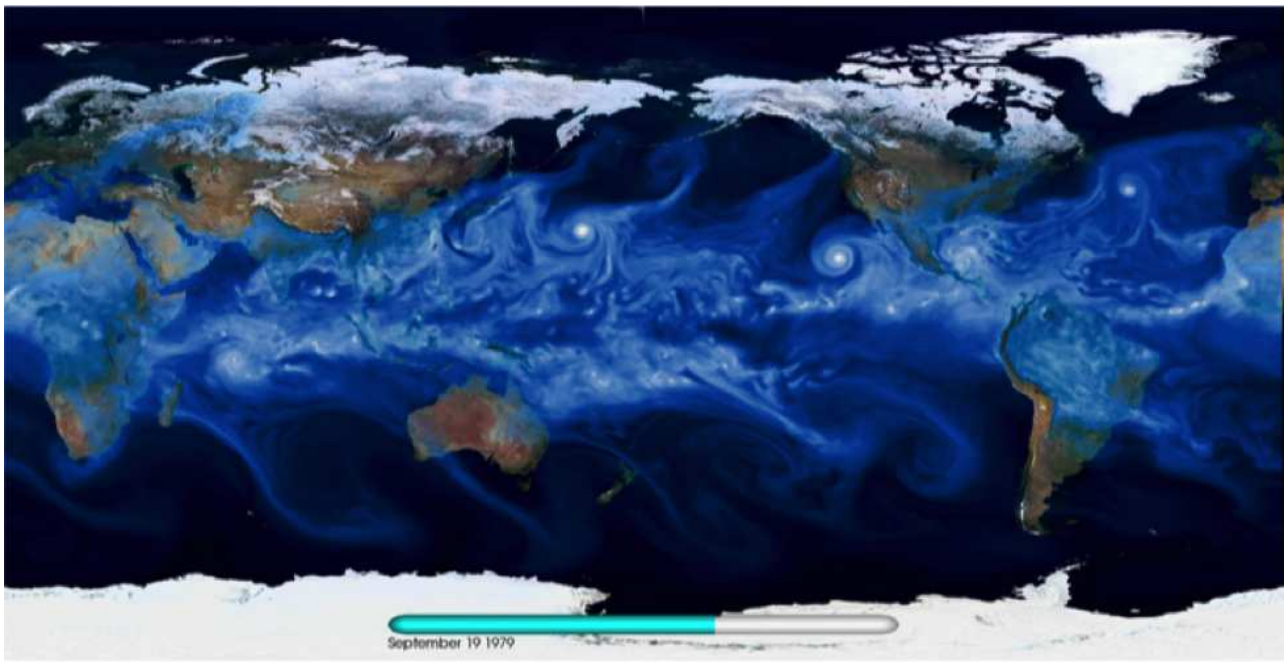

Fig. 3.2: Snapshot of CAM5.1 25-km global climate simulation

cyclones. The study was published in the Journal for Advances in Modeling the Earth System.

"I've been calling this a golden age for high-resolution climate modeling because these supercomputers are enabling us to do gee-whiz science in a way we haven't been able to do before," said Wehner, who was also a lead author for the recent Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC). "These kinds of calculations have gone from basically intractable to heroic to now doable."

Using version 5.1 of the Community Atmospheric Model, developed by the DOE and the National Science Foundation for use by the scientific community, Wehner and his co-authors conducted an analysis for the period 1979 to 2005 at three spatial resolutions: 25 km, 100 km and 200 km. They then compared those results to each other and to observations. One simulation generated 100 terabytes of data. Wehner ran the simulations on NERSC's Hopper supercomputer; the CAM code was optimized for parallel execution and scaling on the Cray system, special emphasis was also laid on the parallel I/O strategy for the code.

"I've literally waited my entire career to be able to do these simulations," Wehner said. The higher resolution was particularly helpful in mountainous areas since the models take an average of the altitude in the grid (25 square km for high resolution, 200 square km for low resolution). With more accurate representation of mountainous terrain, the higher resolution model is better able to simulate snow and rain in those regions.

"High resolution gives us the ability to look at intense weather like hurricanes," said Kevin Reed, a researcher at the National Center for Atmospheric Research and a co-author on the paper. "It also gives us the ability to look at things locally at much higher fidelity. Simulations are much more realistic at any given place, especially if that place has a lot of topography."

The high-resolution model produced stronger storms and more of them, which was closer to the actual observations for most seasons. "In the low-resolution models, hurricanes were far too infrequent," Wehner said. The IPCC chapter on long-term climate change projections concluded that a warming world will cause some areas to be drier and others to see more rainfall, snow and storms. Extremely heavy precipitation was projected to become even more extreme in a warmer world. "I have no doubt that is true," Wehner said. "However, knowing it will increase is one thing, but having confidence about how much and where as a function of location requires the models do a better job of replicating observations than they have."

Wehner says the high-resolution models will help scientists to better understand how climate change will affect extreme storms. His next project is to run the model for a future-case scenario. Further down the line, Wehner believes scientists will be running climate models with 1 km resolution. To do that, they

Fig. 3.3: Big Data Analytics on CMIP-5 data facilitated by ESnet and leadership computing resources at NERSC and ALCF

will have to have a better understanding of how clouds behave.

"A cloud system-resolved model can reduce one of the greatest uncertainties in climate models, by improving the way we treat clouds," Wehner said. "That will be a paradigm shift in climate modeling. We're at a shift now, but that is the next one coming."

In a related exercise, Michael Wehner teamed up with Prabhat (NERSC), Suren Byna (CRD) and Venkat Vishwanath (ALCF) to process the massive CMIP-5 archive at scale on Mira, ALCF's flagship BG/Q system. The team downloaded over 60 TB of climate data from a world-wide repository using the Earth System Grid Federation, pre-processed the data on NERSC's Hopper system, and then transferred 6 TB of data over ESnet to ALCF in 2 days. Prabhat then developed and scaled the TECA (Toolkit for Extreme Climate Analytics) framework, to run on 750,000 cores on ALCF's Mira system. The entire CMIP-5 archive was processed in 1 hour and produced a summary of the expected change in extra-tropical cyclones in future climate change scenarios. It is estimated that a similar task on standalone workstations would take over a decade. This is one of DOE's leading examples of what Scientific Big Data Analytics can accomplish on HPC resources.

Changes in the seasonality of Indian monsoon, availability of fresh water supply through snowpacks in Himalayas and rising sea levels are some examples of the regional impact of global climate change. Characterization of climate change, and adaptation to a changing weather will be key issues for the Indian economy in the 21st century. High resolution simulations through HPC resources, and the application of Big Data Analytics methods on the resulting massive datasets will be key for the scientific community to inform policymakers.

The Life and Death of Stars and the Evolution of the Universe

In recent years, astronomy and cosmology have been transformed from data-starved endeavors into data-intensive sciences. Three key factors have propelled this revolution. First is exponential growth in detector resolution, sensitivity, scale, and reliability. Second is the proliferation of remote, semi-robotic, and fully-robotic telescope operations enabled by powerful networks and intelligent machine scheduling. Third is the fusion of HPC, large-scale databases, and parallel file systems for real-time data analysis and large-scale simulation of astrophysical phenomena.

Today's high-impact astronomical surveys routinely generate >100 GB of digital sky images per night, transfer those images to HPC centers and use automated software pipelines to process the data. These data are then used to generate catalogs of hundreds of millions of objects, and identify time-varying phenomena minutes or hours after they have been observed. Scientists uses these products to identify phenomena for more intensive study using more specialized instruments on the largest telescopes in the world.

The future landscape of astronomy is dominated by the Large Synoptic Survey Telescope (LSST), a facility being constructed in Chile that will generate upwards of 100 PB of data over its lifetime starting in 2022. One of the major surveys in operation today that is paving the way to LSST is the Intermediate Palomar Transient Factory (iPTF, PI Shrinivas Kulkarni, Caltech). The iPTF is an example of how leveraging high-performance networking and computing resources like ESnet and NERSC opens vast new territory in our understanding of the Universe, in this case in the physics of stellar death (supernovae). Understanding supernovae is important because they test our theories of the behavior of matter under extreme conditions, they create and disperse chemical elements heavier than helium, and are useful as tools for measuring distances to study the fundamental physics of Dark Energy. But making progress in this space requires maximizing both data velocity and volume, as iPTF has successfully demonstrated time and again.

For example, iPTF scientists were the first to demonstrate that Type IIb supernovae arise from a kind of massive star called a Wolf-Rayet star. This was the first direct confirmation of the theory, even though the Type IIb supernova phenomenon was first identified some two decades ago. Researchers at Israel's Weizmann Institute of Science were able to identify supernova SN 2013cu within hours of its explosion using the iPTF pipeline running at NERSC. Mere hours after photons from the cataclysmic explosion reached Earth, iPTF was able to trigger

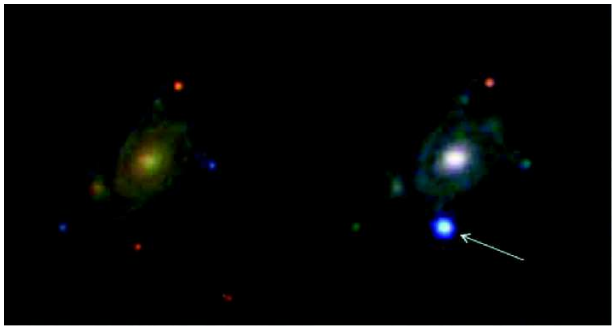

Fig. 3.4: A star in a distant galaxy explodes as a supernova: While observing a galaxy known as UGC 9379 (left; image from the Sloan Digital Sky Survey; SDSS) located about 360 million light-years away from Earth, the team discovered a new source of bright blue light (right, marked with an arrow; image from the 60 inch robotic telescope at Palomar Observatory). This very hot, young supernova marked the explosive death of a massive star in that distant galaxy. Images: Avishay Gal-Yam, *et al.*, Weizmann Institute of Science

telescopes both on the ground and in space to follow the evolution of the supernova more intensively at all wavelengths. These follow-up observations enabled iPTF scientists to determine what elements were present on the surface of the star and in its immediate environment prior to explosion (the findings appeared in the May 22, 2014 edition of *Nature*). The ability to make such discoveries depends on time-critical processing of large volumes of data with HPC, and the ability to identify patterns in the data for scientists to exploit to make new discoveries.

HPC resources not only help scientists transform massive amounts of raw astronomical image data into new knowledge about the life-cycle of stars, it can result in Nobel Prize worth science. In the 1990's the Supernova Cosmology Project (SCP, PI Saul Perlmutter, LBL) used observations of distant supernovae to map out the expansion history of the Universe. Instead of finding that the expansion of the Universe was slowing down, they found that it was speeding up. To eliminate potential sources of systematic error, simulations of the SCP supernova survey were undertaken at NERSC and confirmed the result. Ultimately this discovery led to a Nobel Prize for Saul Perlmutter. This combination of computational science and cosmology led to other projects and established LBL and NERSC as key players in the emerging field of observational cosmology.

There is only one sky, but astronomers are looking deeper into the Universe, opening up new regimes of the electromagnetic spectrum, and examining changes on the timescales of minutes and seconds. "All the low-hanging fruit has been picked by the previous generation of astronomers," notes Berkeley Lab data scientist and astrophysicist Rollin Thomas, "HPC and Big Data are new and essential ladders that lift us up to reach the highest branches."

Key Takeaways

1. Setting up a national resource in HPC and Big Data will require major, sustained investments in hardware. It is recommended that India not embark on the race for flops, but rather focus on well-balanced systems that emphasize compute, memory, storage and networking.

2. In conjunction with investments in hardware, special consideration should be given to system software and applications. *Productivity* of the scientific user community is key, hence investing in purchasing and developing software, and more generally being in sync with the broader open source community is highly recommended.

3. Finally, in our experience, the quality of operational and research staff at such centers is fundamental to the eventual success of such initiatives. Staff needs to be highly qualified, motivated and collaborative; they also need to be compensated appropriately.

Democratization of Data*

Modern science is inherently collaborative, and collaborations produce ever more data. Many large-scale instruments being planned and built that will serve tens of thousands of scientists. These facilities will create petabyte-scale data sets to be analyzed and archived, in many cases using distant computational resources. Though it might seem logical and efficient to house these centers close to their data repositories and computational facilities, this is not

*This section is based on the network research and operational knowledge in ESnet, Energy Sciences Network especially the science requirement reviews, Science DMZ and wide-area data movement.

always the likely scenario. Distributed solutions — in which components are scattered geographically — are much more common at this scale, for a variety of reasons; the largest collaborations will likely depend on distributed architectures.

The LHC, the most well-known high-energy physics collaboration, was a driving force in the development and adoption of such advanced network services. Early on, the LHC community understood the challenges the experiment would present in terms of data generation, distribution, and analysis. In response, the community pioneered a tiered data-distribution model that enables tens of thousands of physicists around the world to access and analyze experimental data. This model is now changing to be more of an 'on-demand' model, where data is moved to the computation, wherever resources are available, and the high-speed networking capabilities are leveraged.

Not just Physics, but many research disciplines are facing the same challenge and marching towards similar solutions. The cost of genomic sequencing is falling dramatically, for example, and consequently, the volume of data produced by sequencers is rising exponentially. In climate science, researchers must analyze observational and simulation data sets located at facilities around the world. Climate data is projected to top 200 petabytes by 2020. The need for productive access to such data led to the development of the Earth System Grid (ESG), a global work-flow infrastructure giving climate scientists access to data sets housed at modeling centers on multiple continents, including North America, Europe, Asia, and Australia.

New detectors being deployed at X-ray synchrotrons generate data at unprecedented resolution and refresh rates. The current generation of instruments can produce 300 or more megabytes per second and the next generation will produce data volumes many times higher; in some cases, data rates will exceed DRAM bandwidth, and data will be preprocessed in real time with dedicated silicon. Large-scale, data-intensive science projects on the drawing board include the International Thermonuclear Experimental Reactor (ITER, the international fusion energy prototype) and the Square Kilometer Array (a massive radio telescope that will generate as much or more data than the LHC).

Role of Networking in Big-data Science

The structure of large-scale science now assumes the availability of high-bandwidth, reliable, feature-rich networks that can interconnect globally-distributed instruments, facilities and collaborators. The Large Hadron Collider at CERN may have been the first experiment for which reliable global networking was a design premise, but it certainly will not be the last.

Within the research and education (R&E) community, instruments, facilities and collaborators are normally served by administratively separate networks. Each of these networks has its own policies, funding models, and technical capabilities. As a result, R&E networking is inherently a multi-domain endeavor. Members of this community spend much of their time coordinating and communicating, in an effort to assure that the global R&E network ecosystem functions optimally from end-to-end.

While this hierarchical, multi-domain, multi-scale model links research facilities no matter where they are located, it is far from seamless. To help address its challenges, various collaborations of R&E networks — including ESnet, Internet2, Regional Optical Networks (RONs), GÉANT, the NRENs of Europe, and other networks from the Americas and Asia — have worked together for decades to develop and standardize technologies and services to assure high performance for scientific data flows from end-to-end.

This infrastructure is only helpful if it can be used effectively by moving data between resources that generate, store and process data. Even if it is within the same supercomputer center, sometimes data movement and data sharing is impeded by architecture and design patterns that are not thought through end-to-end. **The democratization of data i.e. data 'for' all, and shared 'by' all, will be extremely critical in ensuring scientific progress of India**.

Key Issues With Data Sharing Over the Network[1,2]

One of the challenges of building shared information systems for scientific data is that funding models make them actually extensions of existing projects, often going back decades, which have embedded logic and

work practices that are highly resistant to change. In this section, we do not talk about the social causes of data hoarding or business drivers in resisting this, but the infrastructural impediments which prevent users from easily sharing their data.

1. **Inadequate infrastructure at the campus or the data-hosting facility**

   Local area networks are usually general-purpose networks that support multiple missions, the first of which is to support the organization's business operations including email, procurement systems, web browsing, and so forth. Second, these general networks must also be built with security that protects financial and personnel data. Meanwhile, these networks are also used for research as scientists depend on this infrastructure to share, store, and analyze data from many different sources. As scientists attempt to run their applications over these general-purpose networks, the result is often poor performance, and with the increase of data set complexity and size, scientists often wait hours, days, or weeks for their data to arrive, often times they give up getting data over the network.

2. **End hosts not optimized for wide-area data transfers**

   Systems used for wide area science data transfers perform far better if they are purpose-built for and dedicated to this function. When the systems are not designed for data transfer, typically there is a mismatch between the network interface speeds of the end-system, say 10 Gbps, and the capability of the wide-area network, say 1 Gbps. This mismatch overwhelms the WAN connection, and causes packet loss and performance issues for the entire site.

3. **Wide-Area networks are not architected for 'zero packet loss' regardless of their bandwidth capabilities**

   The Transmission Control Protocol (TCP) of the TCP/IP protocol suite is the primary transport protocol used for the reliable transfer of data between applications. TCP is used for email, web browsing, and similar applications. Most science applications are also built on TCP, so it is important that the networks are able to work with these applications (and TCP) to optimize the network for science.

   TCP is robust in many respect—in particular it has sophisticated capabilities for providing reliable data delivery in the face of packet loss, network outages, and network congestion. However, the very mechanisms that make TCP so reliable also make it perform poorly when network conditions are not ideal. In particular, TCP interprets packet loss as network congestion, and reduces its sending rate when loss is detected. In practice, even a tiny amount of packet loss is enough to dramatically reduce TCP performance, and thus increase the overall data transfer time. When applied to large tasks, this can mean the difference between a scientist completing a transfer in days rather than hours or minutes. Therefore, care must be taken when designing networks, with attempts to make it loss-free, so that TCP-based data-intensive science applications perform ideally.

4. **Establishment of a trust-model**

   Data transfer between Science DMZ works within the DOE context since government funding mandates sharing of data at least between other DOE funded scientists and facilities. The centers leverage the grid model to establish trust between the end-systems or a common trusted data movement provider like Globus. In order to facilitate data sharing in the Indian context, it is important to establish the drivers and the trust model, so data sharing is not impeded by issues of trust and verification – such mechanisms must be established by the facilities that host the data and the funding

[1]Dart E, Rotman L, Tierney B, Hester M and Zurawski J "The Science DMZ: A network design pattern for data-intensive science," in *High Performance Computing, Networking, Storage and Analysis (SC), 2013 International Conference for*, vol., no., pp.1-10, 17-22 Nov. 2013

[2]Meyer, E.T. "Moving from small science to big science: Social and organizational impediments to large scale data sharing", *In Jankowski, N. (Ed.), e-Research: Transformation in Scholarly Practice (Routledge Advances in Research Methods series*). New York: Routledge, pp. 147-159, 2009.

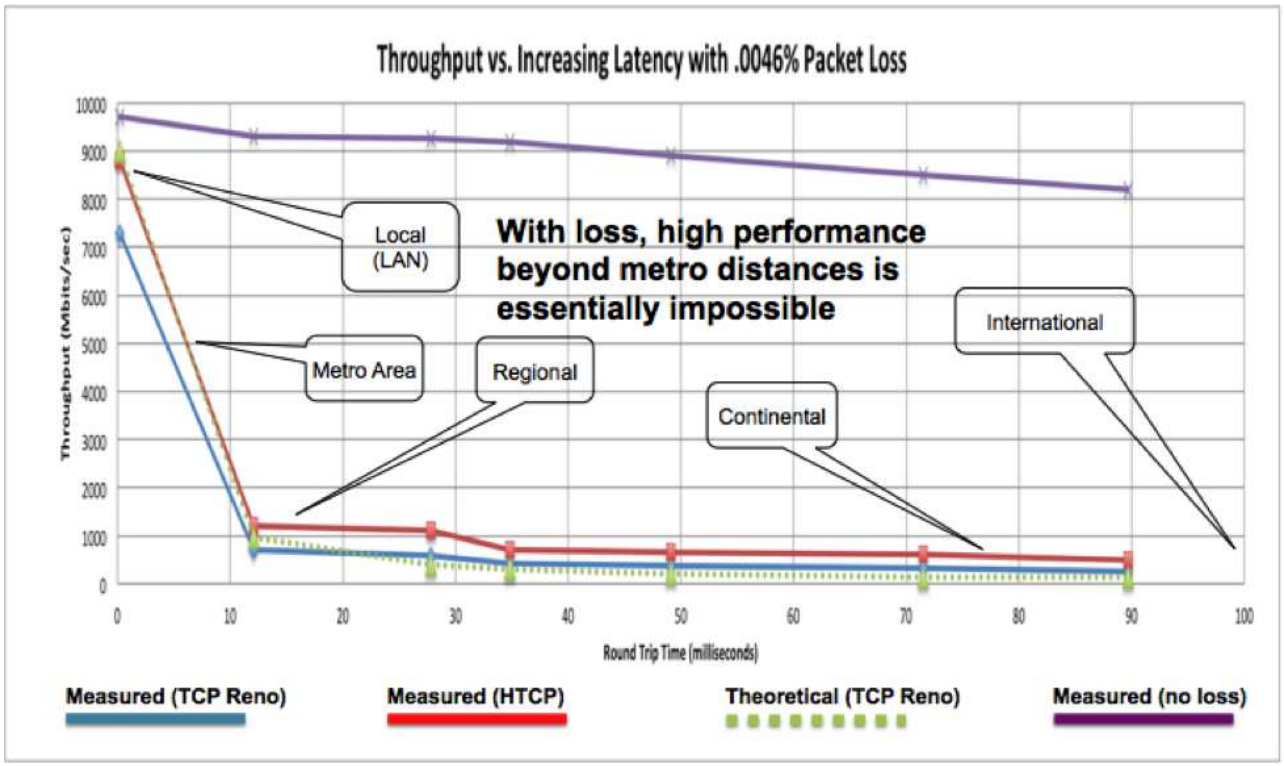

Fig. 3.5: TCP performance as a function of throughput for different sized science networks (note that the performance degradation is quite unique to heavily granular science flows as opposed to classical telco traffic)

organizations that fund research that produce the data-sets.

Science DMZ Infrastructure[1]

As we discussed above, networks optimized for business operations are neither designed for nor capable of supporting the data movement requirements of data intensive science. When scientists attempt to run data intensive applications over these so called "general purpose'' networks, the result is often poor performance — in many cases poor enough that the science mission is significantly impacted and/or the data is shared among the many researchers.

Since many aspects of the campus networks are impossible to change in order to improve performance for everyone, an architecture must be adopted to allow the networks to support science applications without needing to change or impact the general purpose campus network.

The Science DMZ[2] pattern accomplishes this by creating an enclave in the campus network that is engineered for science applications. By separating the data-intensive portion of the network from the general purpose network, it can be assured that the science users get optimal performance to conduct their research while the general-purpose network can be tailored to meet its own purpose.

Scientific collaboration, like any other network-enabled endeavor, is inherently end-to-end. The Science DMZ can easily incorporate wide area science support services, including virtual circuits and software defined networking, and new technologies such as 100 Gbps Ethernet. Developed by ESnet engineers, the Science DMZ model addresses common network performance problems encountered at research institutions by creating an environment that is tailored to the needs of high performance science applications, including high-volume bulk data

[1]For detailed information: http://fasterdata.es.net/science-dmz/

[2]Dart, E.; Rotman, L.; Tierney, B.; Hester, M.; Zurawski, J., "The Science DMZ: A network design pattern for data-intensive science," in *High Performance Computing, Networking, Storage and Analysis (SC), 2013 International Conference for*, vol., no., pp.1-10, 17-22 Nov. 2013

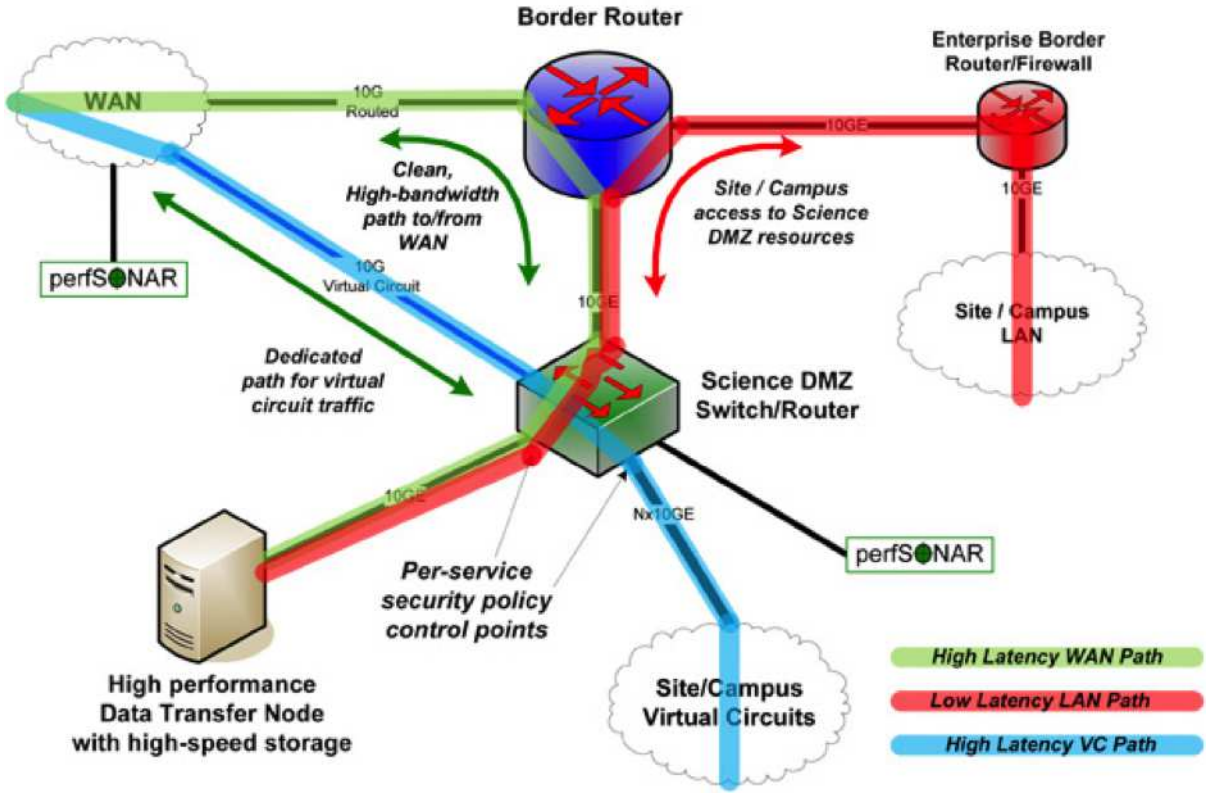

Fig. 3.6: A basic Science DMZ pattern

transfer, remote experiment control, and data visualization.

A Science DMZ design pattern integrates four key concepts that together serve as a foundation for this model. These include:

- A network architecture explicitly designed for high-performance applications, where the science network is distinct from the general-purpose network
- The use of well-tuned, dedicated systems for data transfer
- Performance measurement and network testing systems that are regularly used to characterize the network and are available for troubleshooting
- Security policies and enforcement mechanisms that are tailored for high performance science environments

Key Takeaways

1. It is not enough to just have bandwidth in the local or wide-area network. Care should be taken on how the network is operated and for science big-data movement, a loss-free network architecture and design should be encouraged. Throughput is an end-to-end quality and providing systems and network that are well matched will enable the research effectiveness of scientists in the modern century.
2. Data throughput is an end-to-end problem, and as such only investments in wide-area network will have minimal impact on the data movement ability of scientific data unless similar architectural improvements are undertaken in the campus network. One approach through which campuses can limit their overhaul of the campus network is to build Science DMZ's, enclaves for big-data science and optimized for wide-area data movement.
3. Policies must be set for researchers to share data freely. Such policies need to be backed with funding for data management platforms at well connected institutions so that the data may be

shared without any infrastructure impediment.

Vision for HPC and Data in India

Modern science relies heavily on theory, experiment, simulation and data-driven discovery. Any long term investment in scientific infrastructure should consider these modalities. In particular, facilitating HPC and data-intensive workloads to run efficiently will be key for future progress.

Key Infrastructure Investments Recommended

a. The Government of India should invest in 2-3 strategic computing centers. These centers should have world-class hardware and software resources. The computing centers should ccommodate both supercomputing workloads, as well as data-intensive workloads. Addressing both strong and weak scaling workloads for HPC is important.

b. Investing in developing and growing manpower resources is vital to the long term success and sustainability of a national computing initiative. Hiring and retaining the best national and international talent should be the top priority of such computing centers. Attempts should be made to establish deep connections with leading academic institutions domestically (IITs, IISc) and internationally. Collaborations with leading industry vendors (Intel, Cray, IBM, HP) is highly recommended to keep the center abreast of latest developments. The centers may choose to create an International Advisory Board to keep them abreast of latest developments, and to seek independent evaluation of progress.

c. During the inception phase, these centers should identify key science partners, and develop close working relationships with the relevant scientific institutions and/or communities. We would recommend that the centers align their mission with national scientific priorities.

d. Policies must be established for researchers and facilities to share scientific experimental and simulation data freely. Such policies need to be backed with funding for data management platforms at well-connected[1] institutions (like the strategic computing centers) so that the data may be shared without any infrastructure impediment.

e. Data throughput is an end-to-end problem, and as such only investments in wide-area network will have minimal impact on the data movement ability of scientific data unless similar architectural improvements are undertaken in the campus network. One approach through which campuses can limit their overhaul of the campus network is to build Science DMZ's, enclaves for big-data science and optimized for wide-area data movement.

f. Care should be taken on how the network is architected and operated for science big-data movement. A loss-free network architecture and design should be encouraged. Since application throughput, end-to-end, is the most important metric – providing compute systems, storage systems and network that are well matched will enable the research effectiveness of scientists in India within this modern century.

[1]By 'well-connected' - we mean good network infrastructure at the campus and sufficient bandwidth connectivity to the wide-area network.

References

Dart E, Rotman L, Tierney B, Hester M and Zurawski J (2013) The Science DMZ: A Network Design Pattern for Data-intensive Science. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC 2013, pages **85** 1-85:10, New York, NY, USA. ACM

ESnet-WPa (2015) ESnet Science Requirement Reviews Reports, https://www.es.net/science-engagement/science-requirements-reviews/requirements-review-reports/

ESnet-WPb (2015) Fasterdata by ESriet. http://fasterdata.es.net/

Meyer E T (2009) Moving from Small Science to Big Science: Social and Organizational Impediments to Large Scale Data Sharing. In *Jankowski N (Ed.), e-Research: Transformation in Scholarly Practice (Routledge Advances in Research Methods Series)*, pages 147-159. Routledge, New York

NERSC (2016) NERSC Annual Reports, https://www.nersc.gov/news-publications/publications-reports/nersc-annual-reports/

NERSC-WP (2017) HPC Requirement Reviews, https://www.nersc.gov/science/ hpc-requirements-reviews/.