

LSN Huge Data Workshop

A COMPUTING, NETWORKING AND DISTRIBUTED
SYSTEMS PERSPECTIVE

The Huge Data Workshop was hosted on Zoom
On April 13-14, 2020

*This is a Large Scale Networking IWG workshop
sponsored by National Science Foundation grant CNS-1747856*



LSN Huge Data Workshop

Goal and Scope

*This workshop intends to bring together **domain scientists, network and systems researchers, and infrastructure providers**, to understand the challenges and requirements of “huge-data” sciences and engineering research needs and explore new paradigms to address the problems associated with processing, storing, and transferring huge data.*

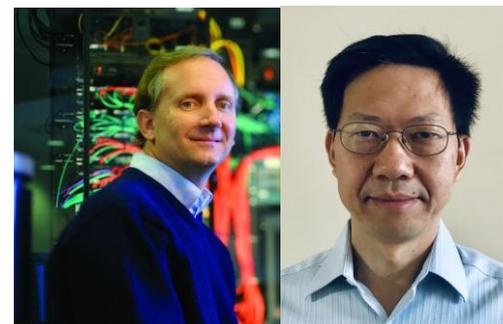
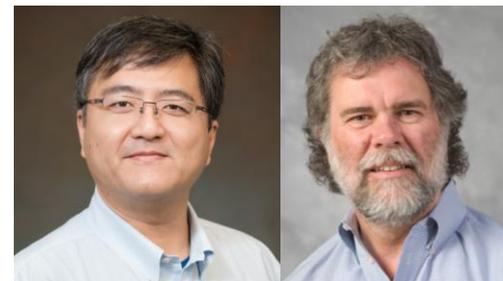
- huge data applications, requirements and challenges
- designing and working with devices for huge data generation
- storage systems for huge data
- software systems and network protocols for huge data
- in-network computing/storage for huge data
- software-defined networking and infrastructure for huge data
- infrastructure support for huge data
- debugging and troubleshooting of huge data infrastructure
- AI/ML technologies for huge data
- measuring the huge data transfer and computation
- scientific workflow of huge data
- access to (portions of) huge data sets
- protecting/securing (portions of) huge data sets



LSN Huge Data Workshop

Organizing Team

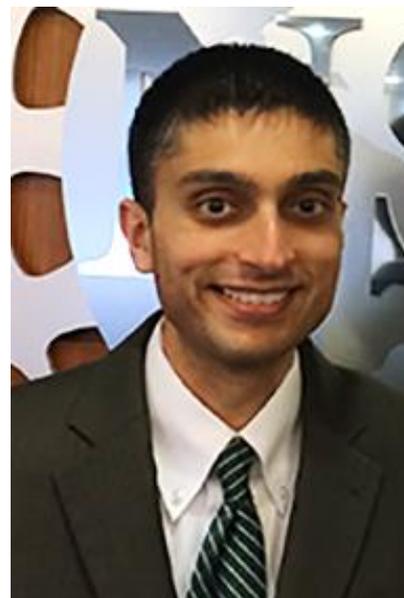
- Organizers
 - Kuang-Ching (KC) Wang, Clemson University
 - Ron Hutchins, University of Virginia
 - Jim Griffioen, University of Kentucky
 - Zongming Fei, University of Kentucky
- Sponsored by NSF CNS #1747856
 - PD: Deep Medhi
- Everything that works today are thanks to
 - Bryttany Todd, RENCi



LSN Huge Data Workshop

NSF Welcome

- Dr. Deep Medhi
CNS Program Director
- Dr. Erwin Gianchandani
CISE Deputy Assistant Director



LSN Huge Data Workshop

Large Scale Networking (LSN) Workshop on Huge Data: A Computing, Networking and Distributed Systems Perspective

Home Program Submission Registration Venue Hotels Upload Slides/Whitepaper

Large Scale Networking (LSN) Workshop on Huge Data: A Computing, Networking and Distributed Systems Perspective

Sponsored by the National Science Foundation (NSF)

Chicago, IL, April 13 – 14, 2020

We have decided to change the workshop to a virtual meeting via zoom.
Online workshop program is now available. Please register for the online workshop by following the Registration link above.

co-located with [FABRIC Community Visioning Workshop](#)

There is an ever-increasing demand in science and engineering, and arguably all areas of research, on the creation, analysis, archival and sharing of extremely large data sets - often referred to as "huge data". For example, the blackhole image comes from 5 petabytes of data collected by the Event Horizon Telescope over a period of 7 days. Scientific instruments such as confocal and multiphoton microscopes generate huge images in the order of 10 GB per image and the total size can grow quickly when the number of images generated increases. The Large Hadron Collider generates 2000 petabytes of data over a typical 12 hour run. These data sets reside at the high end of the "big data" spectrum and can include data sets that are continuously growing without bounds. They are often collected from distributed devices (e.g., sensors), potentially processed on-site or at distributed clouds, and can be intentionally placed/duplicated in distributed sites for reliability, scalability and/or availability reasons. Data creation resulting from measurement, generation, and transformation over distributed locations is stressing the contemporary computing paradigm. Efficient processing, persistent availability and timely delivery (especially over wide-area) of huge data have become critically important to the success of scientific research.

While distributed systems and networking research has well explored the fundamental challenges and solution space for a broad spectrum of distributed computing models operating on large data sets, the sheer size of the data in question today has well surpassed that assumed in prior research. To-date, the majority of computing systems and applications operate based on clear delineation of data movement and data computing. Data is moved from one or more data stores to a computing system, and then it is computed "locally" on that system. This paradigm consumes significant storage capacity at each computing system to hold the transferred data and data generated by the computation, as well as significant time for data transfer before and after the networks, with high performance data transfer functions more closely integrated in software (e.g., operating systems) and hardware infrastructure than have been so far. Such a new paradigm has the potential to avoid bottlenecks for scientific discoveries and engineering innovations through much faster, efficient, and scalable computation across a globally distributed, highly interconnected and vast collection of data and computation infrastructure.

This workshop intends to bring together domain scientists, network and systems researchers, and infrastructure providers, to understand the challenges and requirements of "huge-data" sciences and engineering research needs and explore new paradigms to address the problems associated with processing, storing, and transferring huge data. Topics of interest include, but are not limited to:

- huge data applications, requirements and challenges
- challenges of designing and working with devices for huge data generation
- storage systems for huge data
- software systems and network protocols for huge data
- in-network computing/storage for huge data
- software-defined networking and infrastructure for huge data
- infrastructure support for huge data
- debugging and troubleshooting of huge data infrastructure
- AI/ML technologies for huge data
- measuring the huge data transfer and computation
- scientific workflow of huge data
- access to (portions of) huge data sets
- protecting/securing (portions of) huge data sets

Submission of White Papers

Individuals interested in attending should submit a 1-2 page white paper that addresses a problem related to huge data transfer and processing. White papers should be submitted as PDF attachments by email to hugedata@netlab.uky.edu no later than **February 15, 2020**.

Registration and Travel Grant

A limited number of travel grants are available for authors of accepted white papers to support attendance at the workshop. Registration and travel grant application information can be found by following "Registration/Travel Grant" tab on the top of this page. The deadline is **March 1, 2020**.

Important Dates

Deadline for submission of white papers:	February 15, 2020
Acceptance notification:	February 25, 2020
Registration and travel grants application:	March 1, 2020
Notification of travel grant approval:	March 7, 2020
Workshop dates	April 13-14, 2020

Organizing Committee

Kuang-Ching Wang, Clemson University
James Griffioen, University of Kentucky
Ronald Hutchins, University of Virginia
Zongming Fei, University of Kentucky

Acknowledgment: The workshop is supported in part by the National Science Foundation (NSF) under grant CNS-1747856 and by NITRD Large Scale Networking (LSN) Interworking Group.

TIGER PRINTS CLEMSON LIBRARIES

Home My Account FAQ Contact Us About

Search

Enter search terms: Search

in this collection

Advanced Search

Notify me via email or RSS

Browse by

All Collections

Authors

Expert Gallery

Discipline

Theses & Dissertations

Selected Works Gallery

Journals

Student Works

Conferences

Open Access Fund Collection

Historic Collections

Useful Links

My Account

Contact Us

Author FAO

Author Rights

Scholarly Publishing Information

Home > Conferences > HUGEDATA

LARGE SCALE NETWORKING (LSN) WORKSHOP ON HUGE DATA: A COMPUTING, NETWORKING AND DISTRIBUTED SYSTEMS PERSPECTIVE

This workshop brings together domain scientists, network and systems researchers, and infrastructure providers, to understand the challenges and requirements of "huge-data" sciences and engineering research needs and explore new paradigms to address the problems associated with processing, storing, and transferring huge data. Topics of interest include, but are not limited to:

huge data applications, requirements and challenges challenges of designing and working with devices for huge data generation storage systems for huge data software systems and network protocols for huge data in-network computing/storage for huge data software-defined networking and infrastructure for huge data infrastructure support for huge data debugging and troubleshooting of huge data infrastructure AI/ML technologies for huge data measuring the huge data transfer and computation scientific workflow of huge data access to (portions of) huge data sets protecting/securing (portions of) huge data sets.

[2020 Program](#)

Browse the contents of Large Scale Networking (LSN) Workshop on Huge Data: A Computing, Networking and Distributed Systems Perspective:

[Introduction](#)

[Data Storage](#)

[Data Processing and Security](#)

[Data Movement](#)

[Data Generation](#)

Reader from: Singapore, Singapore, Singapore

5G, Edge Computing and Future Network Support for Huge Data (Presentation)

Zhi-Li Zhang



Recent Downloads

20 of 51 in the past week

76 Total Papers	458 Total Downloads	448 Downloads in the past year
-----------------	---------------------	--------------------------------

Embed

Terms of Use

Close

We use cookies to help provide and enhance our service and tailor content. By closing this message, you agree to the [use of cookies](#).

LSN Huge Data Workshop

Keynote

- Craig Partridge, Colorado State U., Internet Hall of Fame
“Are Our Networks Trashing Our Files?”
- Huge data + Fast network + Fast wireless + Big storage = ?
 - Packet size exceeds CRC-32 error detection capability
 - Middleboxes known to rewrite checksum for corrupted data
 - Link level error rates increase as wireless gets faster
 - !! Lots of huge data with errors in distributed storage goes undetected and is driving our sciences !!



LSN Huge Data Workshop

Lightening Talks

<https://docs.google.com/spreadsheets/d/1dpEWz4SE8AMJF9jiQTH5yIJ5dWR4lepGY9JN7B37WDs/edit?usp=sharing>

- Day 1
 - Data generation (6 talks)
 - Data Storage (5 talks)
- Day 2
 - Data movement (14 talks)
 - Data processing & security (14 talks)



LSN Huge Data Workshop

- Huge data problems span: data generation, movement, storage, computing
- Huge data is often due to continuous generation, too big, and often not meaningful to store forever
- Huge data problems tend to be domain-specific
 - Brian imaging, Astronomy, Pathology, Radiology, Genomics, Connected & automated vehicles, IoT, High energy physics, Antenna, ...
 - Exabytes per year, continued growth, distributed globally
- Systems research is key – architecture, integration
- Huge data is breaking current infrastructure – errors!
- Need new infrastructure addressing huge data centric challenges
 - Scale, error tolerance, new methods, application-centric, plan ahead



LSN Huge Data Workshop

Breakouts

- Breakout 1 – New Areas of Research
- Breakout 2 – New Types of Data
- Breakout 3 – Cross-disciplinary Collaboration
- Breakout 4 – Critical infrastructure



LSN Huge Data Workshop

“Key Issues” Discussion on Day 1

- 67 participants
- Key issues brought up
 - Errors in data transport protocols (e.g., TCP, QUIC, ...)
 - Silent corruptions, rotten bits ended up in data stores (see keynote)
 - Need stronger error checking mechanisms to detect silent corruptions
 - How to navigate huge data without downloading the whole
 - Need new methods – access, compute
 - Commercial cloud data cost models not scalable – data ingress/egress fees
 - For huge data that is real-time/near-real-time
 - Use it or lose it
 - Streaming data computation, “peek”/search
 - Data lakes for universities



LSN Huge Data Workshop

“Key Issues” Discussion, Contd.

- Key issues, continued:
 - Revisit networking solutions – expose underlying resources, e.g., network buffers, programmability based on traffic
 - Security vs. storing sensitive data in shared data repository, e.g., PHI
 - Ongoing efforts on data transfer tools and training - PRP, NSRC, Globus, ...
 - This is a systems problem – Need to integrate disparate solutions and optimize
 - What is the right data-centric architecture? E.g., NDN is focusing on data policy, caching, working on transporting LHC data
 - Distributed storage – federated access is easy. Performance hit! Use SSD!
 - Middleboxes, CDNs, Firewalls
 - Not all users are using the best (data transfer, parallel computing) tools!



LSN Huge Data Workshop

“Key Observations” Discussion on Day 2

- 62 participants
- Key observations:
 - Huge data challenges are multi-facet
 - Large size, static and streaming data, across multi-domains
 - Altogether breaks the current paradigm
 - Huge data problems tend to be application specific
 - Not just needing more infrastructure, but overwhelming today’s technologies
 - Increasing need of edge computing
 - Increasing multi-use for research and production, e.g., city/campus data
 - Real-time, use it or lose it data – much of the huge data is useless after some period of time
 - Data Challenges
 - Provenance, metadata, naming, FAIR principles
 - Closer integration of networking and processing
 - Data generation is not a separate problem, dep. on application, network, processing, storage
 - Lossy compression
 - Data quality, lifecycle



LSN Huge Data Workshop

“Key Observations” Discussion, Contd.

- Key observations, continued:
 - Big takeaways
 - Huge data challenges are seen across sciences
 - It is about a conscious choice of content, tradeoff data volume vs. computation – we don't need all the data
 - Huge data breaks paradigms – need new systems to fix where things are broken
 - Even light processing data at the edge and/or in the network is valuable
 - Integration of science, applications, and infrastructure is more important than ever
 - The network must be data aware – capable of processing, storing, delivering data
 - Today's cloud model is not sustainable for huge data paradigm – computing is cheap, storage is expensive, networking is even more expensive.



"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development
(NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674,
Fax: 202-459-9673, Email: nco@nitrd.gov, Website: <https://www.nitrd.gov>

