Northeastern

SANDIE: Named Data Networking for Data Intensive Science

Edmund Yeh, Harvey Newman, Christos Papadopoulos

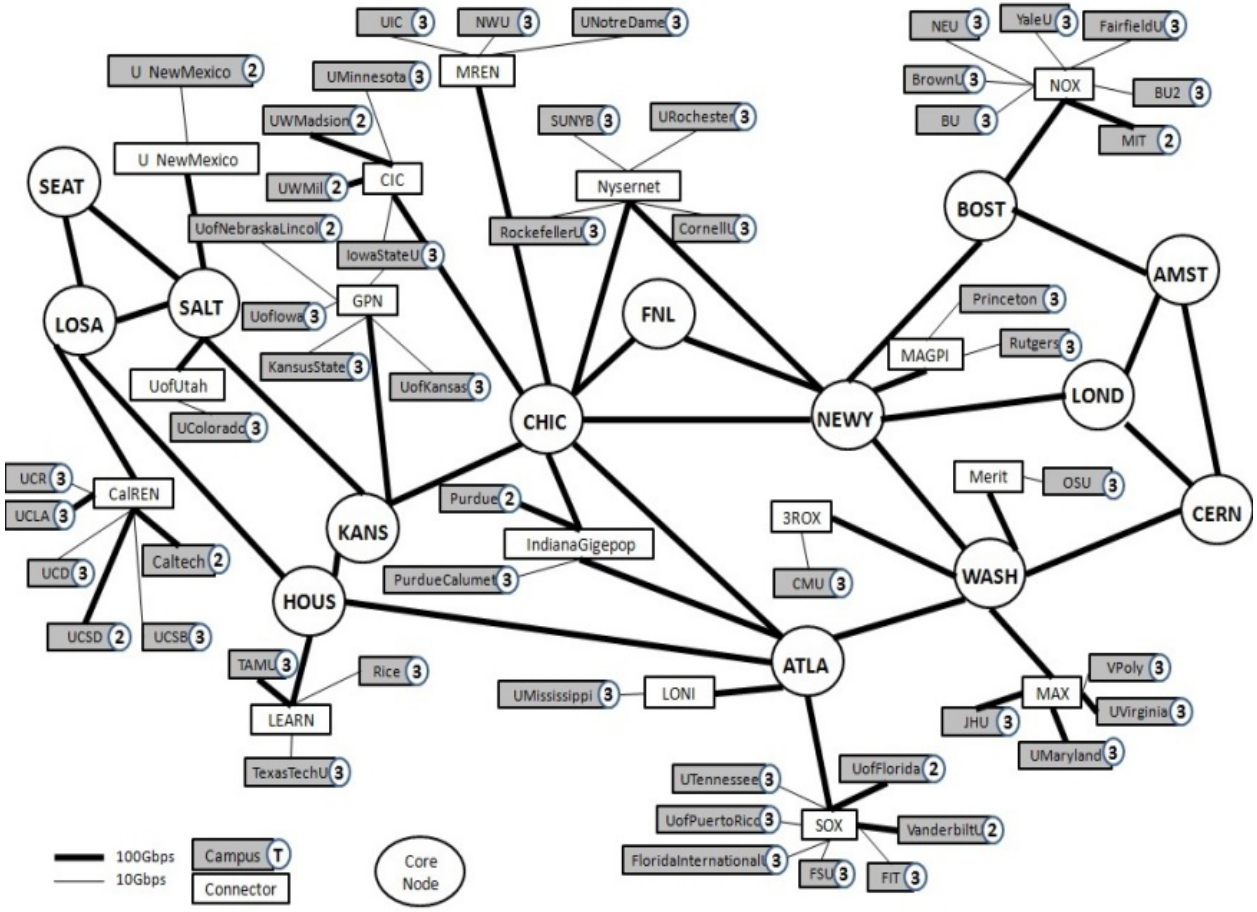
Data-intensive Science and LHC

- Data-intensive science: extract knowledge from massive datasets with growing scale and complexity.
- Global data distribution, processing, access, analysis; coordinated use of large but limited computing, storage and network resources.
- **NDN a natural fit for data-intensive science.**
- Large Hadron Collider (LHC) high energy physics program is **world's largest data intensive application**, handles about 1 Exabyte (1 million terabytes) by 2018.
- LHC network connects CERN to 500 tiered sites worldwide.
- Used to make Nobel-prize discoveries such as Higgs Boson.
- LHC network traffic projected to grow by another 2 orders of magnitude by 2026.
- Increased complexity of data.
- Present system cannot scale to meet needs of HC Run3 (2021-23).

SANDIE Project

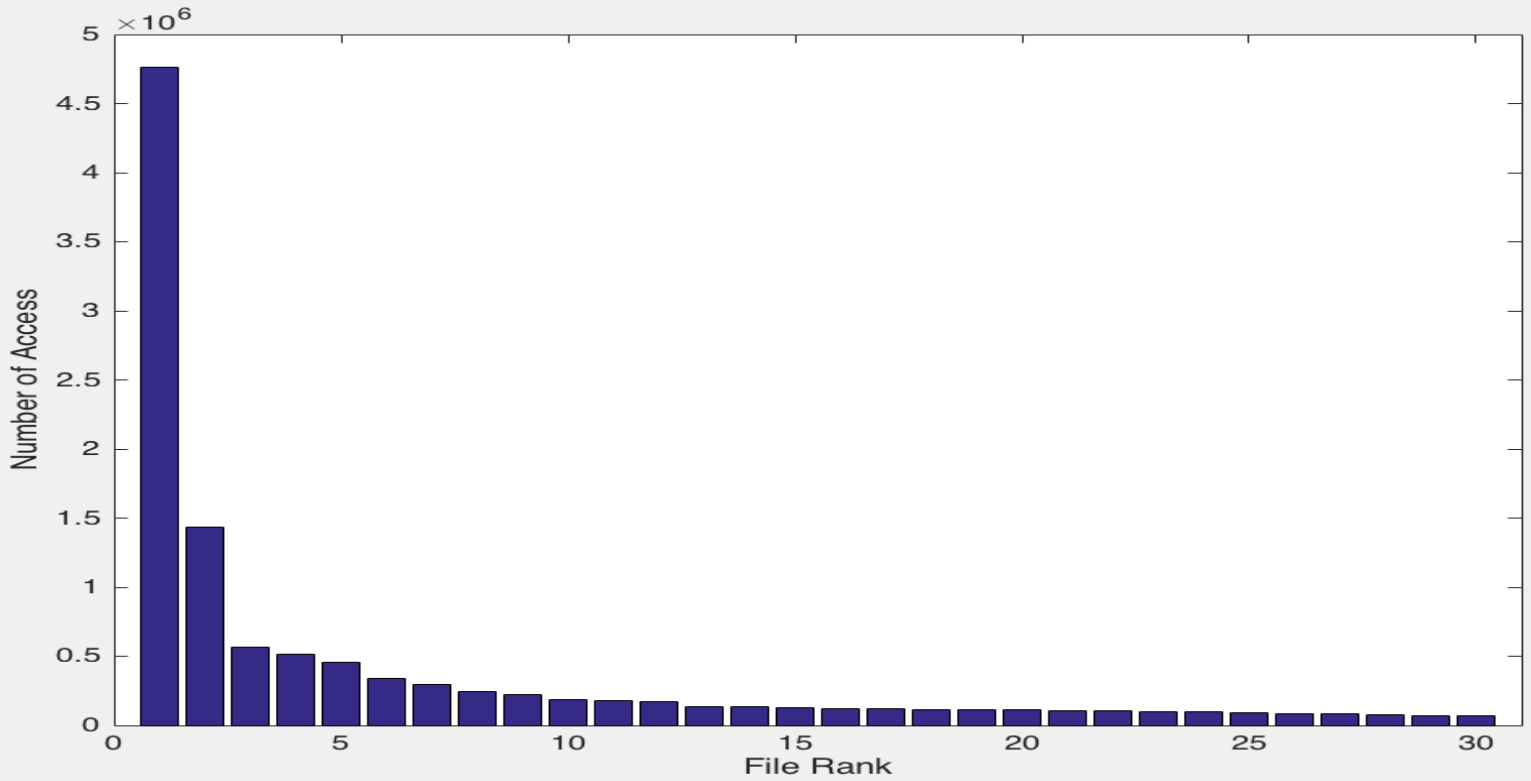
- $1M, 2-year US NSF CC* grant starting July 2017.
- Northeastern (lead), Caltech, Colorado State.
- Use NDN principles to redesign high energy physics (HEP) Large Hadron Collider (LHC) program network.
- Will deploy 10 NDN edge caches with SSDs and 40G/100G network interfaces at 7 sites (Northeastern, Caltech, CSU, Fermilab, ..)
- Combine with larger core caches in strategic locations.
- Coordinate with compact preprocessing of data objects (e.g. MiniAOD, MicroAOD, data scouting, n-tuples).
- Leverages NDN-based tested for climate applications at CSU.

LHC Network Map



Northeastern

LHC Data Access Patterns



Files	Total Size	Min Size	Max Size	Mean Size	Median Size
58035992	149.86 PB	0.00 MB	163.82 GB	2.58 GB	2.59 GB

Project Aims

- Derive NDN-based operational model for data distribution, processing, gathering, analysis to benefit LHC and other major data-intensive scientific programs (e.g. LSST, SKA).
- Simultaneous optimization of caching (“hot” datasets), forwarding, and congestion control in network core and site edges; leverages PI previous work.
- Network core and edge node design and performance optimization.
- Development of naming scheme and attributes for fast access and efficient communication in HEP and other fields.
- Scalability of NDN system in amount and type of data supported as well as geographic extent.