Made possible by grants from the NSF and the Schmidt Foundation

Open Storage Network Seminar Series

Vision

Alex Szalay, JHU

January 21, 2021

NSF grants
1747552
1747493
1747507
1747490
1747483

National Storage Infrastructure

- Storage infrastructure largely balkanized
  - Every campus/project does its own specific vertical system
  - As a result, lots of incompatibilities and inefficiencies
  - People are only interested in building minimally adequate
  - As a result, we build storage tiers ‘over and over’
  - Big projects need petabytes, also lots of ‘long tail’ data
- Cloud storage not a good match at this point for PBs
  - Amazon, Google, Azure too expensive:
    you have to buy the storage HW every month
  - Wrong tradeoffs: cloud redundancies too strong for science
  - Getting data out is very expensive
  - But: this will change over time

Everybody needs a reliable, industrial strength storage tier!

OpenStorageNetwork.org

- **NSF CC*: 150+ universities to connected at 40-100G**
- Ideal for a large national distributed storage system:
  - Inexpensive (~$120K) storage (1.5PB) racks at each site (~200PB)
  - Store much of the NSF generated data
  - Provide significant shared storage for Big Data Hub communities
  - Distribute data from MREFC projects
  - Provide gateways/caches to XSEDE and cloud providers
- Technology straightforward
  - Automatic compatibility, ultra-simple standard API (S3)
  - Implement a set of simple policies
  - Enable sites to add additional storage blocks at their own cost
  - Variety of services built on top by the community
- Estimated Cost: ~$20M for 100 nodes

* Current partnership:
  $1.8M NSF,
  $1.0M Schmidt Foundation
* Built a 6 node testbed and demonstrated feasibility
* Establish wide community support, through the Big Data Hubs

SCHMIDT FAMILY FOUNDATION

NSF

Rapidly establish the third main pillar for NSF science infrastructure

COMPUTE | **NETWORKING** | **DATA**

Potential Impact

- Totally change the landscape for academic Big Data
  - Create a homogeneous, uniform storage tier for science
  - Liberate communities to focus on analytics and preservation
  - Amplify the NSF investment in networking
  - Very rapidly spread best practices nationwide
  - Universities can start thinking about PB-scale projects
- Impact unimaginable
  - Links to XSEDE
  - Big Data projects can use it for data distribution
    - LHC, LSST, OOI, genomics
  - Small projects can build on existing infrastructure
  - Enable a whole ecosystem of services to flourish on top
  - Would provide “meat” for the Big Data Hub communities
    - E.g. enable nation-wide smart cities movement

New opportunity for federal, local, industrial, private partnership