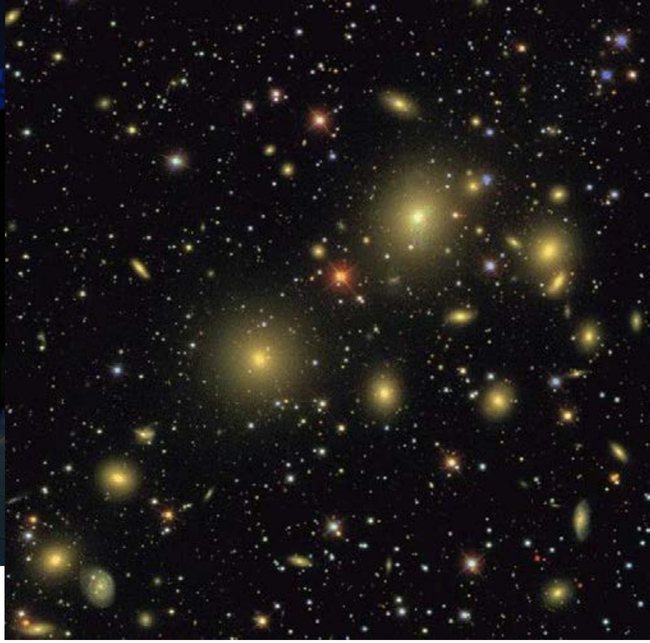




Building the Open Storage Network

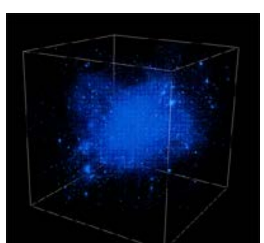
Alex Szalay
The Johns Hopkins University

Institute for Data Intensive Engineering and Science
idies

Emerging Trends in Science

- Broad sociological changes
  - *Convergence of Physical and Life Sciences*
  - *Data collection in ever larger collaborations*
  - *Virtual Observatories: CERN, IVOA, NCBI, NEON, OOI,…*
  - *Analysis decoupled, off archived data by smaller groups*
- Scientific data sets moving from 100TBs to PBs
  - *While the data are here, analysis solutions are not*
  - *Data preservation and curation needs to be reinvented*
- National infrastructure doesn't map onto new needs

SDSS



BIOLOGY PHYSICS ETHICS INFORMATICS ENGINEERING CHEMISTRY

Computational Infrastructure

- The NSF has invested significant funds into high performance computing, both capacity and capability
  - *These systems form XSEDE, a national scale organization with excellent support infrastructure*
  - *The usage of these machines is quite broad, and gradually transitioning from HPC simulations to include more and more large data analysis tasks*
- Most large MREFC projects still build their own computational infrastructure in a vertical fashion

Current Storage Landscape

- Storage 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:*
    they force you to buy the storage every month
  - *Wrong tradeoffs: cloud redundancies too strong for science*
  - *Getting data in (and out) is very expensive*

Everybody needs a reliable, industrial strength storage tier

Opportunity

- The NSF has funded 150+ universities to connect to Internet2 at high speeds (40-100G) for ~$150M
- Ideal for a large national distributed storage system:
  - *Place a 1-2PB storage rack at each of these sites (~200PB)*
  - *Create a redundant interconnected storage substrate using an industrial strength erasure code storage*
  - *Incredible aggregate bandwidth, easy flow between the sites*
  - *Can also act as gateways to cloud providers*
  - *Automatic compatibility, simple standard API (S3)*
  - *Implement a set of simple policies*
  - *Enable sites to add additional storage at their own cost*
  - *Variety of services built on top by the community*
- Estimated Cost: $30-40M

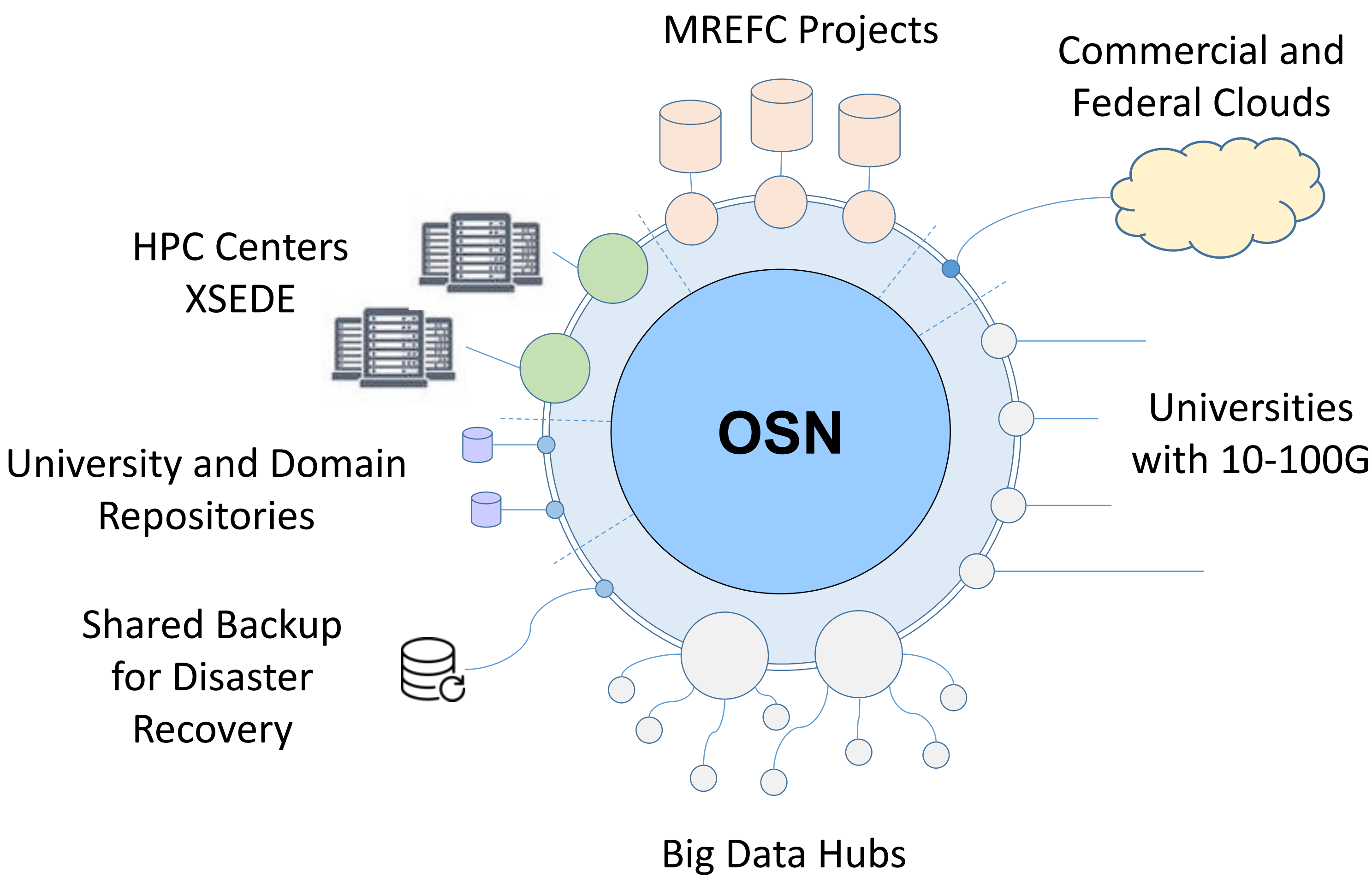
System could be the world's largest academic storage facility

Transformative 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, NDS, RDA, Globus*
  - *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*
    - Enable nation-wide smart cities movement

New opportunity for federal, local, industrial, private partnership

Connections



Questions, Tradeoffs

Cannot do "everything for everybody"!

- Where to draw the line? Use the 80-20 rule…
  - *Build the 20% of possible, that serves 80% of needs*
- Hierarchical or flat?
  - *A single central 'science cloud' vs a totally flat ring?*
  - *Or 4-6 big sites with 10-20PB, the rest flat with 1-2PB?*
- Object-store or POSIX
  - *Keep it simple, focus on large objects*
- This is really a social engineering challenge
  - *Teach the universities how to be comfortable with PB data*
  - *Centralized may be more efficient, but will have trust issues*
  - *Giving each university its own device speeds up adaptation*

High-Level Architecture

- Should there be any computing on top?
  - *A lightweight analytics tier makes system much more usable*
  - *A set of virtual machines for front ends*
  - *But these also add complexity?*
  - *Everybody needs similar storage, analytics tier more diverse*
  - *Some need HPC, others Beowulf/ Hadoop/ TensorFlow/ ??*

- Focus on simplicity
  - *Everybody needs storage*
  - *Create a simple appliance with 1-2PB of storage*
  - *100G interfaces, straddling the campus firewall and DMZ*
  - *Ultra simple object-store interface, possibly S3*
  - *Maybe built in Globus Lite*

Building Blocks

- Scalable element (SE)
  - *500TB of storage+ single server*
  - *Support 40G interface for sequential read/write*
  - *Should saturate 40G for read, about half for write*
- Stack of multiple SEs
  - *Aggregated to 100G on a fast TOR switch, now becoming quite inexpensive (<$20K)*
- These can also exist inside the university firewall
  - *But purchased on local funds, storing local data*
- Software stack to be discussed
  - *ZFS, Ceph, Mero,…*
  - *Integrated with Globus "Lite", with streamlined stack*

Management

- Who owns it?
  - *OSN storage should remain in a common namespace*
  - *This would enable uniform policies and interfaces*
- Software management
  - *Central management of software stack (push)*
  - *Central monitoring of system state*
- Hardware management
  - *Local management of disk health*
  - *Universities should provide management personnel*
- Policy management
  - *This is **hard** and requires a lot more discussion*
- Monitoring
  - *Two tier, store all events and logs locally, send only alerts up*
  - *Try to predict disk failures, preventive maintainance*
- Establish metrics for success

Security Ideas

- How do we make sure the system is secure?
  - *Appliances exist in DMZ*
  - *IPSEC across nodes?*
- How do we connect through the university firewalls?
  - *Possibly a second interface inside firewall, access is subject to the university authentication*
  - *Only push/pull from the inside*
- Need lots more input from security experts

The Road Towards OSN

1. Establish public / private partnership
   - *Early seed founds from the Eric Schmidt Foundation*
2. Build community prototypes for different use cases
   i. *Move and process 1PB of satellite images to Blue Waters*
   ii. *Move specific PB-scale MREFC data from Tier1 to Tier2 at a university for detailed sub-domain analytics (LSST)*
   iii. *Create large simulation (cosmology or CFD) at XSEDE and move to a university to include in a NumLab*
   iv. *Take a large set of LongTail data with small files and organize into larger containers, and explore usage models*
   v. *Interface to cloud providers (ingress/ egress/ compute)*
3. Build community initiative for large scale funding

Summary

- High end computing has three underlying pillars
  - *Many-core computing/HPC / supercomputers*
  - *High Sped Networking*
  - *Reliable and fast data storage*
- The science community has heavily invested in first 2
  - *Supercomputer centers/XSEDE, Internet 2, CC-NIE, CC**
- Time for a coherent, national scale solution for data
  - *Needs to be distributed for wide buy-in and **TRUST***
- Only happens if the whole community gets behind it