# Use of S3 storage

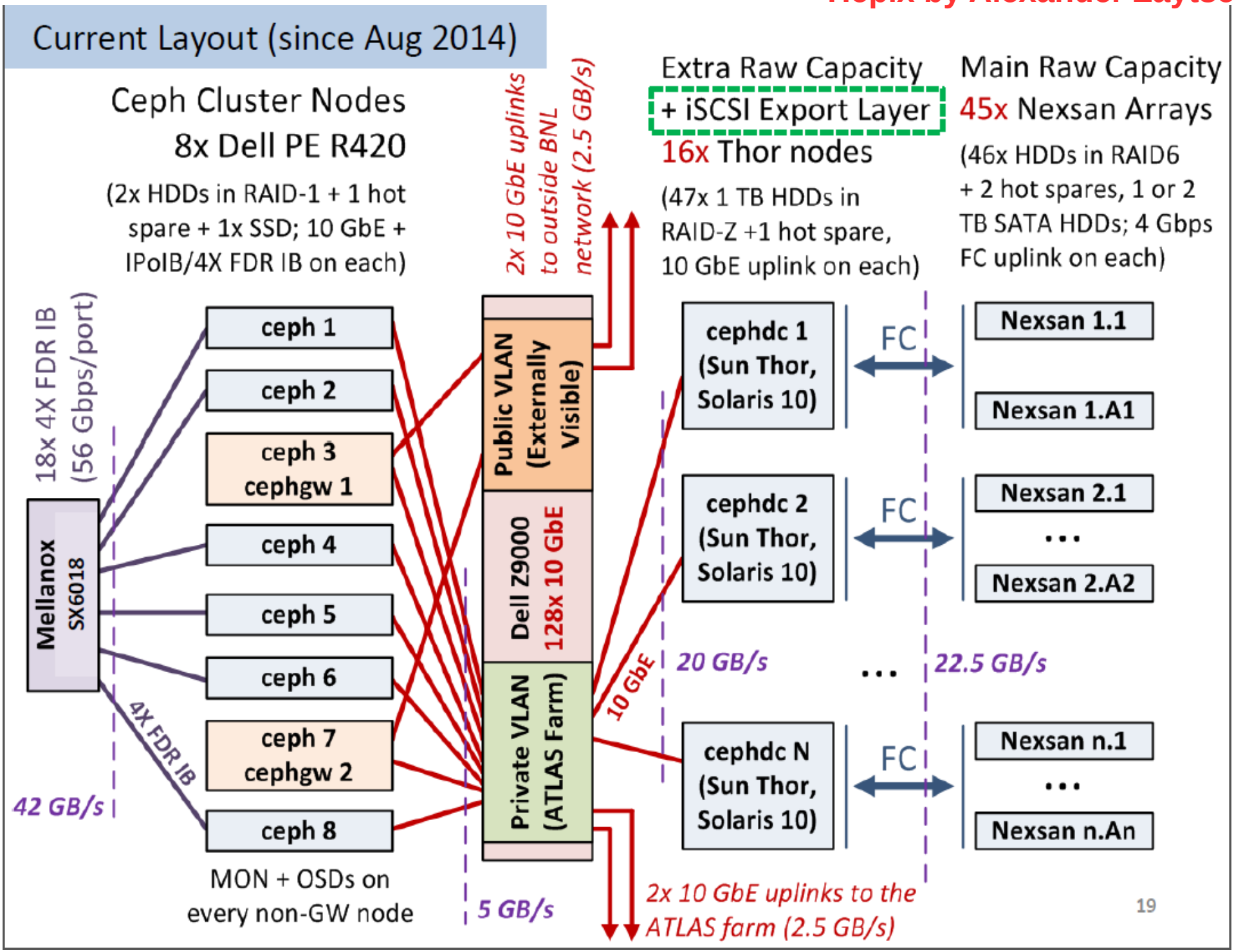
Hironori Ito
Brookhaven National Laboratory

# S3 Storage

- S3 is a storage API for various object storages.
- Amazon, Ceph, Riak, xRootd, etc... are supporting S3 interface.
- Within the S3 storage, one stores objects(files/data) in a bucket(directory/place-holder).
- One can store large number of objects in a bucket or create large number of buckets.
- There is generally no associated file system.
  - Fuse is possible.
- To access S3 storage, one needs access_key_id and secret_access_key.
  - access_key_id is associated with buckets and their permissions.

# BNL Ceph

- BNL has currently RAW 1.8 PB of storage using the retired storage systems.  With the replication factor of 3, the actual usable space is about 0.6PB.
- The Ceph cluster consists of 8 head nodes.  Two of them are also used as the access gateway via S3 API as well as regular http.
  - The detail is shown in the next slide by Alexander Zaytsev (also see his HEPIX presentation http://indico.cern.ch/event/320819/session/6/contribution/39 )
- BNL is currently in the process to increase the performace and capacity by using more retired storages.
- BNL, MWT2 and AGLT2 are also dicussing the possible test of Federated Ceph storage.
  - It is one-master + many slaves style storage.
  - http://ceph.com/docs/master/radosgw/federated-config/


Current Layout (since Aug 2014)
Ceph Cluster Nodes
8x Dell PE R420
(2x HDDs in RAID-1 + 1 hot spare + 1x SSD; 10 GbE + IPoIB/4X FDR IB on each)
18x 4X FDR IB (56 Gbps/port)
Mellanox SX6018
ceph 1
ceph 2
ceph 3
cephgw 1
ceph 4
ceph 5
ceph 6
ceph 7
cephgw 2
ceph 8
4X FDR IB
42 GB/s
MON + OSDs on every non-GW node
2x 10 GbE uplinks to outside BNL network (2.5 GB/s)
Public VLAN (Externally Visible)
Dell Z9000
128x 10 GbE
Private VLAN (ATLAS Farm)
5 GB/s
Extra Raw Capacity
+ iSCSI Export Layer
16x Thor nodes
(47x 1 TB HDDs in RAID-Z +1 hot spare, 10 GbE uplink on each)
Main Raw Capacity
45x Nexsan Arrays
(46x HDDs in RAID6 + 2 hot spares, 1 or 2 TB SATA HDDs; 4 Gbps FC uplink on each)
cephdc 1 (Sun Thor, Solaris 10)
cephdc 2 (Sun Thor, Solaris 10)
cephdc N (Sun Thor, Solaris 10)
FC
Nexsan 1.1
Nexsan 1.A1
Nexsan 2.1
Nexsan 2.A2
Nexsan n.1
Nexsan n.An
10 GbE
20 GB/s
22.5 GB/s
2x 10 GbE uplinks to the ATLAS farm (2.5 GB/s)
19

# BNL's Ceph Use

- PANDA developers have been using BNL Ceph storage for outputs of job logs at relatively small scale.
  - Pilots developers (Paul and Wen) have added S3 features using boto python API.
  - PANDA server developers (Tadashi) has added the supports for storage access_key_id and secret_access_key
- PANDA developers have also asked/suggested BNL to create the simple http interface to access the logs.
  - Ceph S3 supports the simple http natively.  However, it can not be used due to our network configurations.
  - Simple RAILs Restful webapp was created as proxy service to the backend storage.
    - It is simply access via http://host/bucket/object
- The storage has been stable without any major issues.

# Amazon S3

- BNL has been ingaging with Amazon to use their cloud services as a possible ATLAS computing and storage resources.
- Amazon has three storages; S3 (simple storage), EBS (elastic block store) and Glacier (backup)
  - http://aws.amazon.com/s3/
- BNL has been testing S3 as a possible, regular ATLAS storage within Amazon.
  - Read/writes by pilots
  - Read/writes by DDM

# Access to Amazon S3

- Pilots use S3 APIs to read/write S3 (just like Ceph)
  - Needs access_key_id and secret_access_key, which are stored in the PANDA server
- DDM
  - DDM currently needs the SRM.
    - By Carlos Gamboa
    - Install BestMan SRM
    - BestMan SRM needs file system
      - Mount the S3 storage as Fuse.
  - End points
    - BNL-AWSEAST_DATADISK/PRODDISK/USERDISK were added to AGIS.
    - DDM transfers has been tested for functionality and performance
      - ~ a few 100 MB/s writing to S3 due to using one stream.
        - Multiple streams breaks checksum validation
      - ~ several 100 MB/s reading from S3.
        - Multiple streams are fine.
      - The performance can be increased.
    - The most recent FTS3 supports direct S3 access without the use of SRM.
      - Needs testing
      - DDM must also supports S3 to utilize FTS3's S3 capabilities.