State of the Lustre File System

Reliability, Resiliency, and Community Roadmap

Mini-Symposium on Data over Distance

Shawn Hall, BP – Shawn.Hall@bp.com
OpenSFS Director at Large

What Is OpenSFS?

- OpenSFS is a vendor neutral, member supported non-profit organization bringing together the open source file system community

- Our mission is to aggregate community resources and be the center of collaborative activities to ensure efficient coordination of technology advancement, development, and education

- The end goal is the continued evolution of robust open source file systems for the community

OpenSFS Reorganization

- Completed in 2017
- Why?
  - Establish total user community control over OpenSFS
  - Broaden the membership base
  - Increase participation from the members
  - Distribute responsibility equally among the members

OpenSFS Structure

- Two levels of membership
  - Members
    - Lustre end users
  - Participants
    - Lustre vendors
- User community controlled and driven
  - Board members can only be selected from the Members
  - Members vote to elect the Board and on changes to OpenSFS
- Flat and low annual membership fee
  - $1000 – Members
  - $5000 – Participants

OpenSFS Responsibilities

- Organize LUG
- Collect feature and development requests from the Members
- Produce an annual document summarizing the requests for Participants
- Provide a unified voice for the user community to Lustre vendors

OpenSFS Board

President: Sarp Oral, ORNL

Vice President: Kevin Harms, ALCF

Treasurer: Kirill Lozinskiy, NERSC

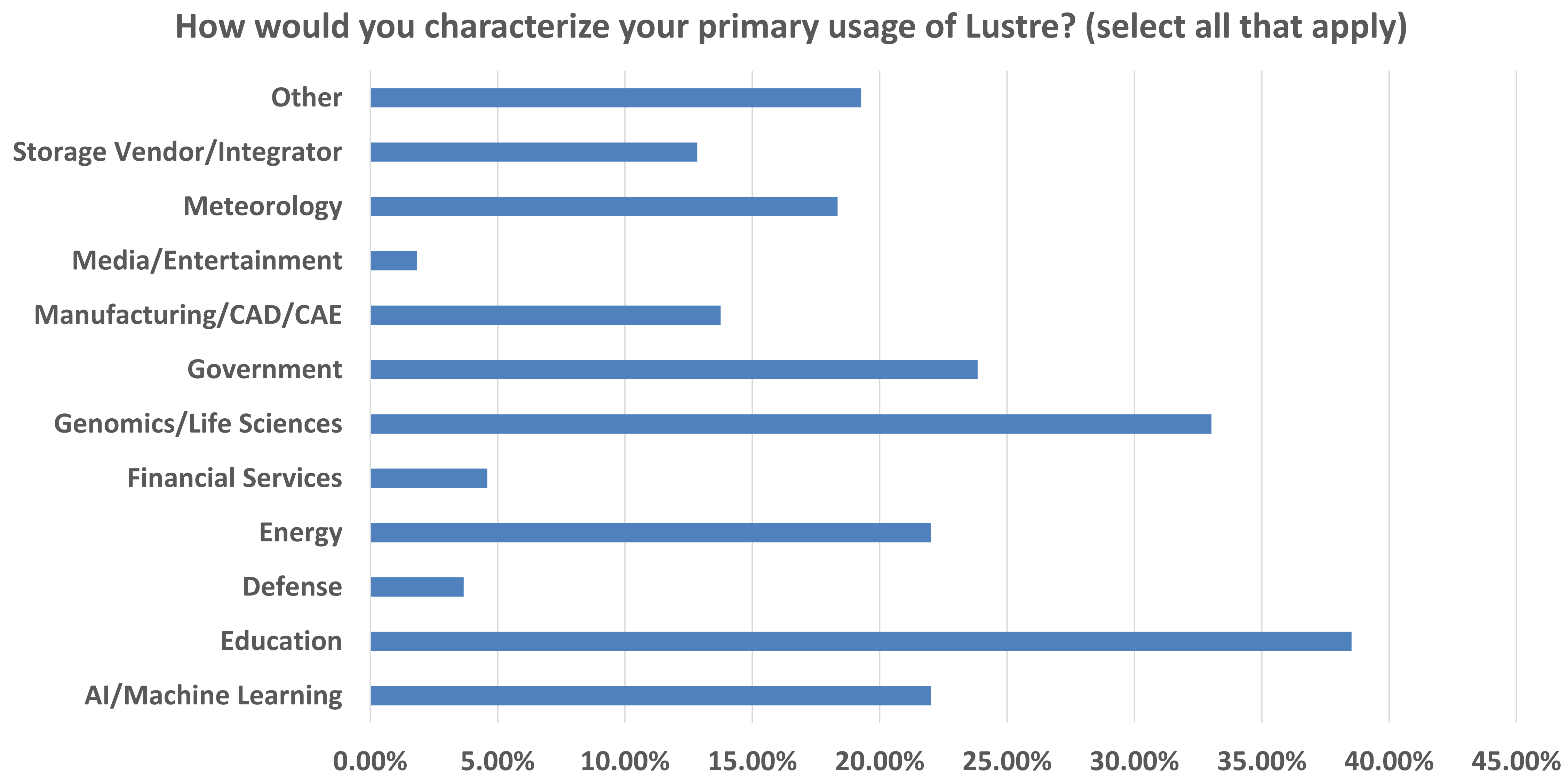
Secretary: Ken Rawlings, IU

Director At Large: Shawn Hall, BP

Lustre on the Move

- Seagate Lustre Engineering business has been acquired by Cray in 2017
- Intel Lustre Engineering business has been acquired by DDN/Whamcloud in 2018
- Development and test infrastructure also relocated
  - Issue tracking https://jira.whamcloud.com
  - Patch reviews https://review.whamcloud.com
  - Download site http://downloads.whamcloud.com
  - Wiki https://wiki.whamcloud.com
- http://opensfs.org/press-releases/opensfs-congratulates-ddn-on-acquiring-intels-lustre-file-system-capability/

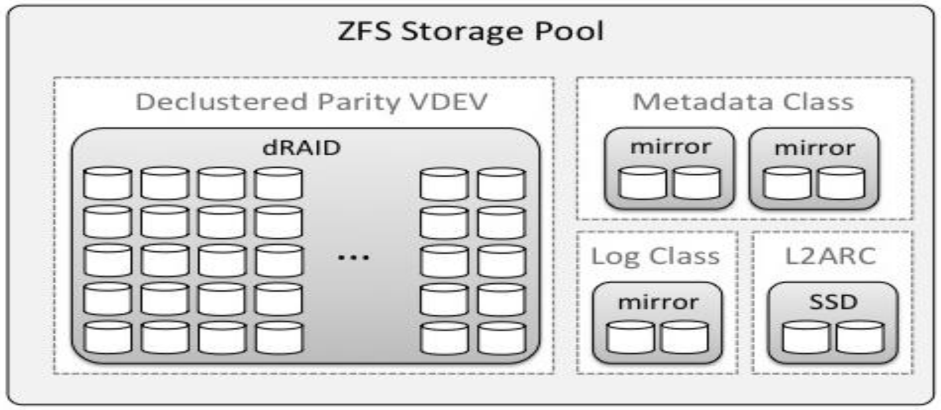
Community Survey - Usage



- AI/Machine Learning much higher than anticipated
- Need to refine categories based on feedback in Other

Lustre – Current State

Lustre on ZFS

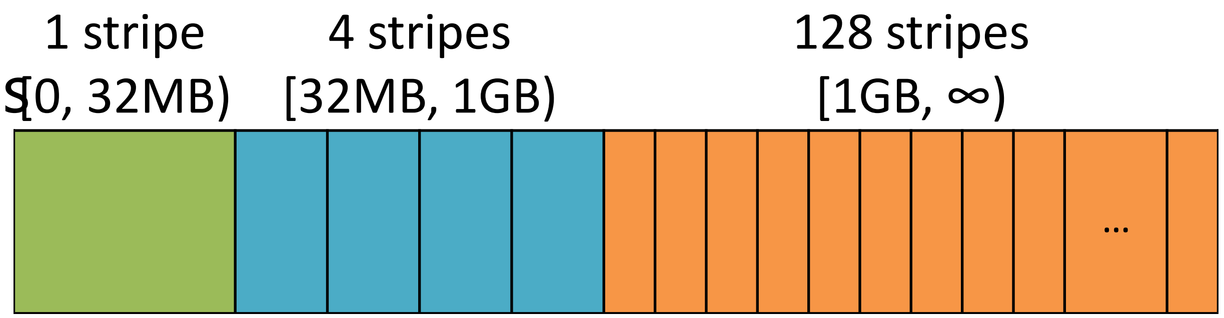


- Changes for using ZFS more efficiently
  - Improved file create performance (Intel
  - Snapshots of whole file system (Intel)
- Changes to core ZFS code
  - inode quota accounting (Intel)
  - Multi-mount protection for safety (LLNL)
  - System and fault monitoring improvements (Intel, HPE)
  - Large dnodes for improved extended attribute performance (LLNL)
  - Reduce CPU usage with hardware-assisted checksums, compression (Intel)
  - Declustered parity & distributed hot spaces to improve resilvering (Intel)
  - Metadata allocation class to store all metadata on SSD/NVRAM *(Intel)*

Composite File Layouts

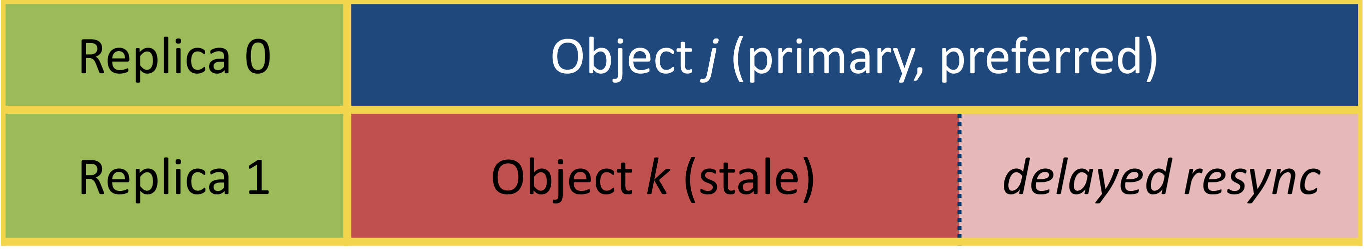
- Composite File Layout allows different layout based on file offset
  - Provides flexible layout infrastructure for upcoming features
    - File Level Redundancy (FLR), Data-on-MDT (DoM), HSM partial restore, etc
  - Layout components can be disjoint (e.g. PFL) or overlapping (e.g. FLR)
- Progressive File Layout (PFL) simplifies usage for users and admins (ORNL)
  - Optimize performance for diverse users/applications
  - One PFL layout could be used for all files
  - Low stat overhead for small files
  - High I/O bandwidth for large files

Example progressive file layout with 3 components



File Level Redundancy

- Provides significant value and functionality for both HPC and Enterprise use
  - Select layout on a per-file/dir basis (e.g. mirror all input data, one daily checkpoint)
  - Higher availability for server/network failure - finally better than HA failover
  - Robustness against data loss/corruption - mirror or M+N erasure coding for stripes
  - Increased read speed for widely shared files - mirror input data across many OSTs
- Replicate/migrate files between storage classes
  - NVRAM->SSD->HDD
  - Local vs. remote replicas



Replica 0

Object *j* (primary, preferred)

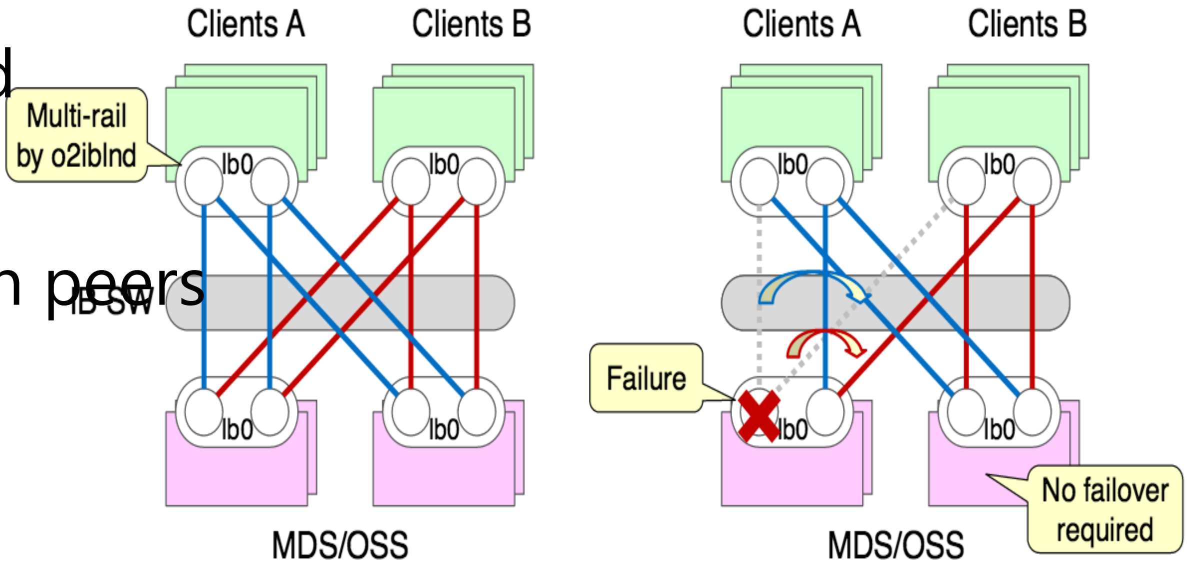
Replica 1

Object *k* (stale)

delayed resync

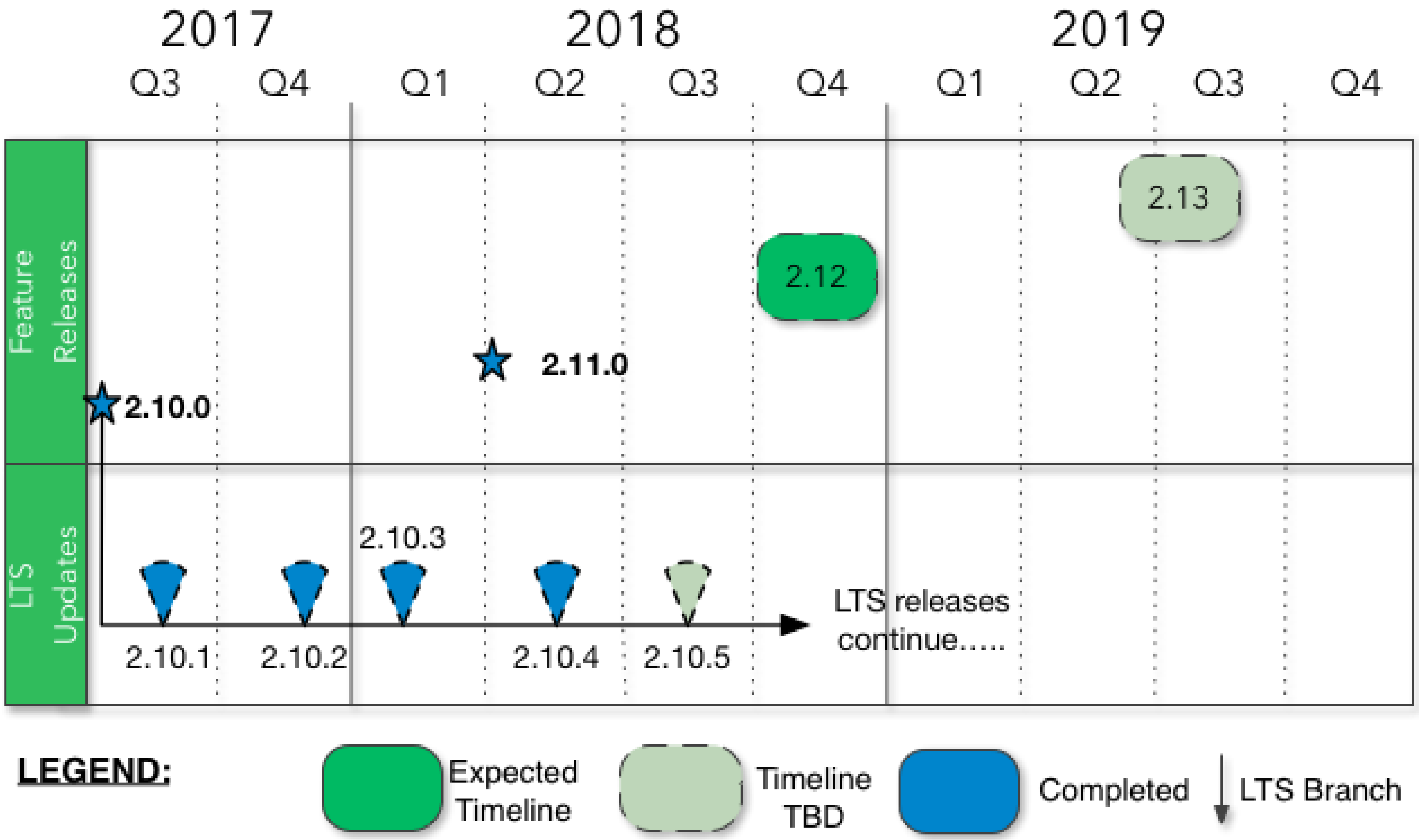
Multi-Rail Lnet

- Allow LNet across multiple network interfaces
  - Supports all LNet networks – LNet layer instead of LND layer
  - Allows concurrent use of different LNDs (e.g. both TCP and IB at one time)
- Scales performance significantly
  - Scaling limits currently being tested
- Improves reliability
  - Active-active network links between peers



Lustre – Future Plans

Lustre Community Roadmap



2.10
- **ZFS Snapshots**
- **Multi-rail LNet**
- **Progressive File Layouts**
- **Project Quotas**

2.11
- **Data on MDT**
- **FLR Delayed Resync**
- **Lock Ahead**

2.12
- **FLR Framework Advancements**
- **LNet Health**
- **DNE Dir Restriping**

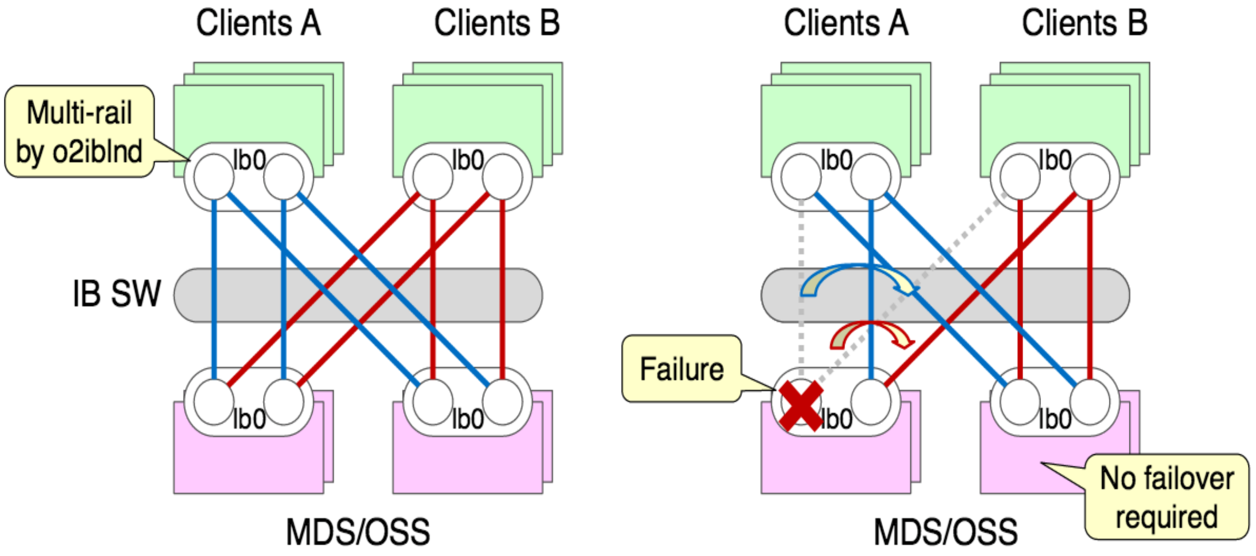
2.13
- **FLR Erasure Coding**
- **Persistent Client Cache**

* Estimates are not commitments and are provided for informational purposes only

* Fuller details of features in development are available at http://wiki.lustre.org/Projects

LNet Network Health, UDSP (2.12/2.13)

- Builds on LNet Multi-Rail in 2.10/2.11 (LU-9120 Intel, HPE/SGI*)
  - Detect network interface and router failures automatically
  - Handle LNet fault w/o lengthy Lustre recovery, optimize resend path
- User Defined Selection Policy (LU-9121 Intel, HPE*)
  - Fine grained control of interface selection
  - Optimize RAM/CPU/PCI data
  - Useful for large NUMA mach



FLR Erasure Coded Files (LU-10911 Intel 2.13)

- Erasure coding gives redundancy without 100% or 200% mirror overhead
- Add erasure coding to new or existing striped files *after* write is finished
  - Use delayed/immediate mirroring for files being actively modified
- Suitable for striped files - add N parity per M data stripes (e.g. 16d+3p)
  - Parity declustering avoids IO bottlenecks, CPU overhead of too many parities
    - e.g. split 128-stripe file into 8x (16 data + 3 parity) with 24 parity stripes

Lustre – Additional Work

Lustre over the WAN

- Analytics of Wide-Area Lustre Throughput Using LNet Routers – Nageswara Rao, ORNL
  - http://wiki.lustre.org/images/4/41/LUG2018-Wide_Area_Throughput_LNet_Routers-Rao.pdf
- Lustre/ZFS at Indiana University – Stephen Simms, IU
  - https://www.ddn.com/download/lustre-zfs-at-indiana-university-2/?wpdmdl=41124&refresh=5b4ffd32aa6fe1531968818
- Lustre nodemap/shared key
  - https://lustre.ornl.gov/ecosystem-2016/documents/tutorials/Simms-IU-NodemapSharedKey.pdf
  - http://cdn.opensfs.org/wp-content/uploads/2015/04/GSS-Shard-key-Update-and-Using-UID_Simms.pdf
  - https://scholarworks.iu.edu/dspace/bitstream/handle/2022/20645/Thota%20et%20al%202016.pdf?sequence=4&isAllowed=y

Thank you

Open Scalable File Systems, Inc.
3855 SW 153rd Drive
Beaverton, OR 97006
Ph: 503-619-0561
Fax: 503-644-6708
admin@opensfs.org

OpenSFS™

www.opensfs.org

Who likes buy one get one free deals?

High Performance Remote Graphics

Shawn Hall

Background

- BP has a single High Performance Computing center in Houston (AKA “the HPC”)
- Researchers globally use the HPC
- Impractical to replicate compute, storage, and networking resources at every major site
- What is easier to move across the network – terabytes of compressed seismic data or compressed pixels?
    - We use Expedat from Data Expedition to move data quickly
    - https://www.dataexpedition.com/
- **Answer: compressed pixels**

Questions from management

- Most researchers already have a nice workstation at their desk, so why don't we just shove all their workstations in the server room close to the data?
  - Cost inefficient
  - Space inefficient
  - Hard to administer
- Can we just use servers and virtualize them? What do we do with the workstations?

Cost comparison

Dell Precision 7820 Workstation

- 1 x Intel Xeon Gold 6140 18C @ 2.3 GHz
- Nvidia Quadro P4000 8 GB
    - Subtract $598 for P400 2 GB
- 96 GB RAM 2666 MHz DDR4
- 2 x 3.5” 1 TB 7200 RPM HDD
- $9,875 (Dell website)
- $945 (remote graphics software)
- **Grand total: $10,820**

Dell R740 Rack Server

- 2 x Intel Xeon Gold 6140 18C @ 2.3 GHz
- Nvidia Tesla P40 24 GB
- 768 GB RAM 2666 MHz DDR4
- 8 x 3.5” 1 TB 7200 RPM HDD
    - Just use NAS w/ dedup in reality
- $51,936 (Dell website)
- $16,380 (8 VMs - software licensing)
- $43,453 (24 VMs – software licensing)
- Sliced into 8 VMs = **$8,540 per user**
    - Oversubscribed 2:1 = **$4,270 per user**
- Sliced into 24 VMs = **$3,975 per user**
    - Oversubscribed 2:1 = **$1,987 per user**

Oversubscription benefits of virtualization

- Oversubscribe CPU cores
  - E.g. Give each VM 18 cores on a 36 core host
- Oversubscribe memory*
  - Memory ballooning add/removes memory from VMs based on pressure
- Oversubscribe GPU
  - vGPU memory is fixed, but vGPU can use all host GPU cores
  - Requires vGPU scheduling policy change**
- Oversubscribe network
  - 40/100 GbE host network shared by VMs

* In theory – has not been successfully tested w/ vGPU
** http://docs.nvidia.com/grid/latest/grid-vgpu-user-guide/index.html#changing-vgpu-scheduling-policy

Components used at BP

- Server with Nvidia Tesla series GPU
  - Tesla GPU required for GPU virtualization
- Red Hat Virtualization
- Nvidia vGPU host and guest drivers
- Mechdyne TGX remote graphics software
  - Captures video of remote system and sends to local system
- Leostream Connection Broker software
  - Acts as traffic officer directing users to available systems

Nvidia vGPU

- Gives greater flexibility for choosing slices of GPU allocated to VMs
- Fewer GPU cards with multiple GPU chips per card makes vGPU a necessity

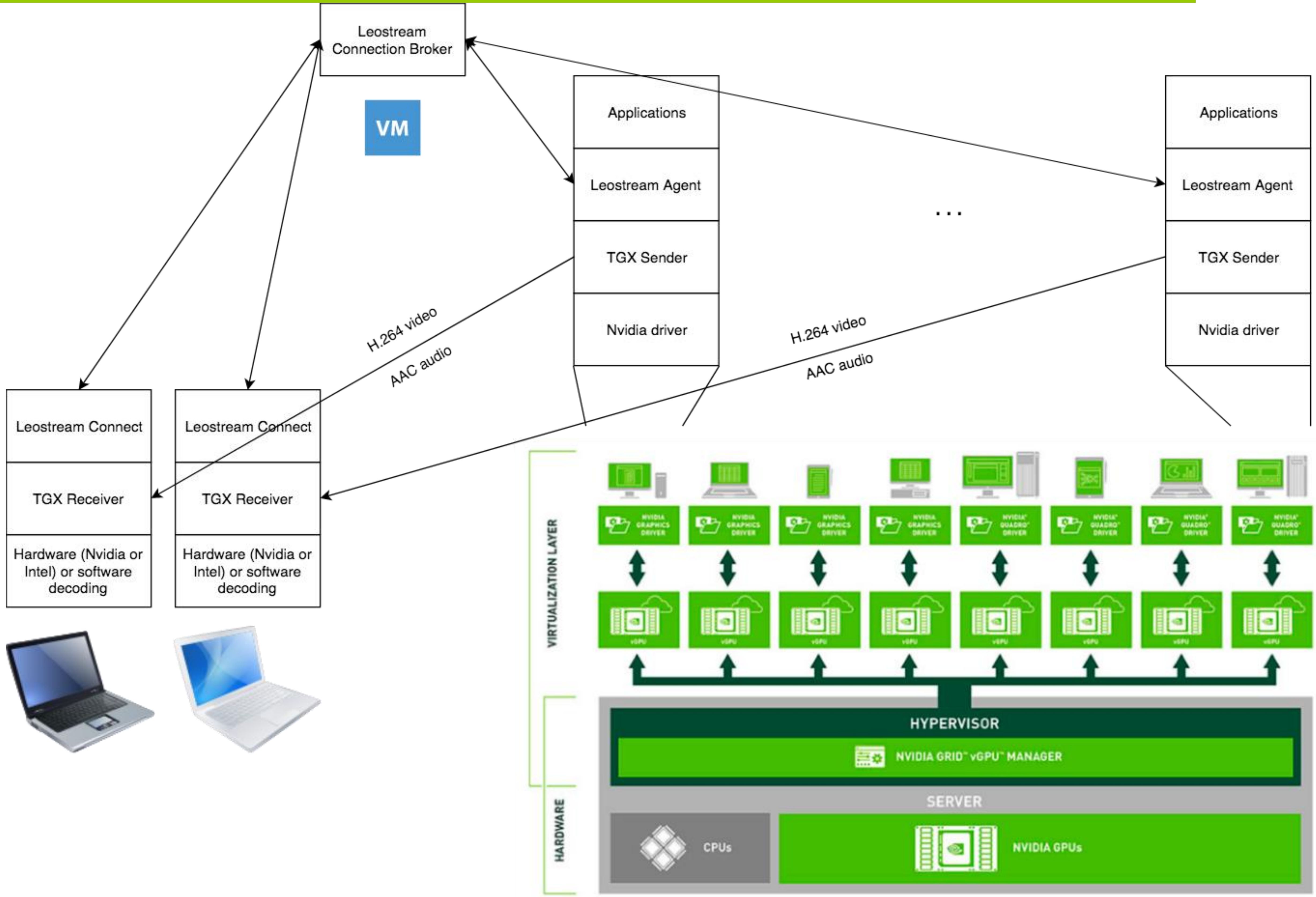
Mechdyne TGX

- Offloads computation through Nvidia Capture SDK to GPU to minimize latency
- High color accuracy
- Multi user collaboration
- Data is SSL encrypted
- Compatible with Windows, Mac, and Linux
- Performance improvements come “automatically” with GPU improvements

Leostream

- Very flexible, vendor agnostic connection broker
- Directs users to available systems based on policies
- Uses 3 components
  - Leostream Connection Broker – routes users to a desktop, manages TGX connection settings, defines policies around desktop usage
  - Leostream Connect – runs on user's computer, establishes and configures TGX session based on Connection Broker response
    - Windows, Mac, Linux clients
  - Leostream Agent – runs on remote computer and helps the Connection Broker to manage connections

Architecture overview



Leostream Connection Broker

VM

Applications

Leostream Agent

TGX Sender

Nvidia driver

Applications

Leostream Agent

TGX Sender

Nvidia driver

. . .

H.264 video

AAC audio

H.264 video

AAC audio

Leostream Connect

TGX Receiver

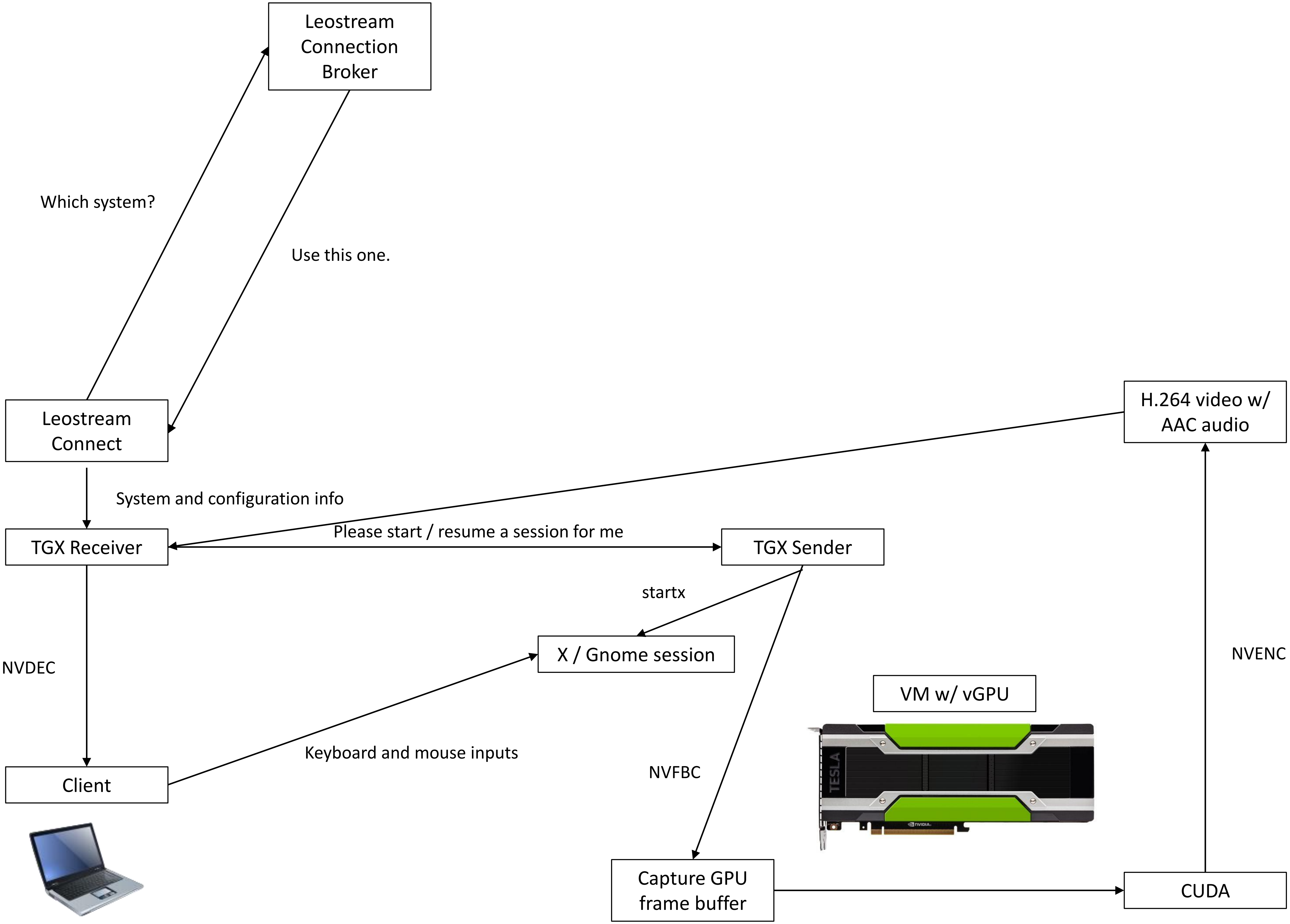
Hardware (Nvidia or Intel) or software decoding

Leostream Connect

TGX Receiver

Hardware (Nvidia or Intel) or software decoding

Architecture detail



Leostream Connection Broker

Which system?

Use this one.

Leostream Connect

System and configuration info

TGX Receiver

Please start / resume a session for me

TGX Sender

H.264 video w/ AAC audio

startx

X / Gnome session

NVDEC

NVENC

VM w/ vGPU

Keyboard and mouse inputs

NVFBC

Client

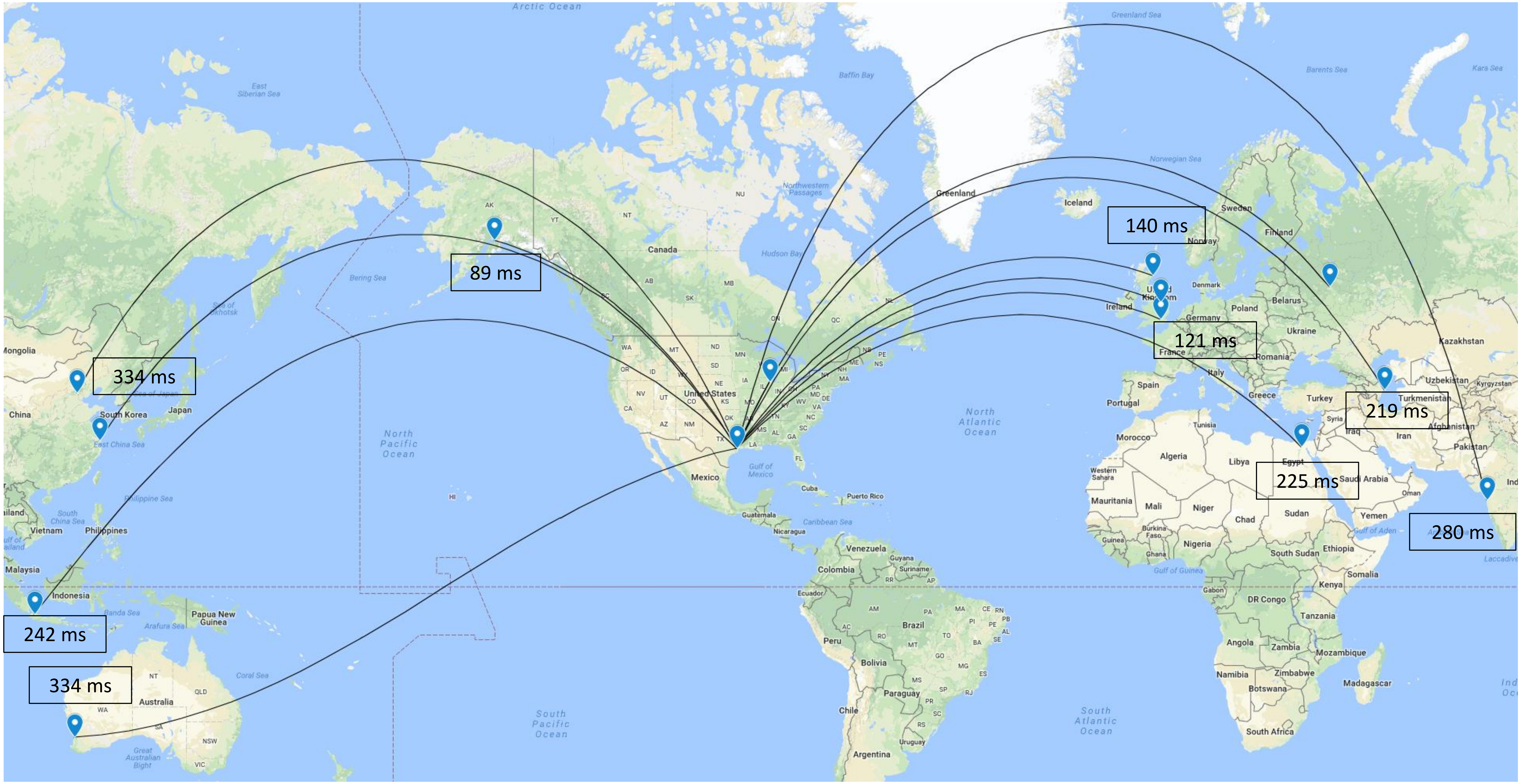
Capture GPU frame buffer

CUDA

System management

- Systems are provisioned with a minimal kickstart file
- System configuration is managed with Ansible
  - CentOS and RHEL VM images
  - Red Hat Virtualization hosts
  - Red Hat Virtualization manager (provisioning VMs)
  - (To-do) health checks and / or cleanup tasks on VMs

Use in BP



Testimonials

- *Previously, I've been very limited in bandwidth here in the UK (often < 200 Kb), as I connect over SSH the StarCCM+ graphics front end to the CFD simulations running on the HPC. X Windows never worked and the connection has been painfully slow, with several minutes lag for several types of operation. TGX has overcome all these problems and is making the process of interacting with HPC more easy and efficient.* – London, UK based CFD engineer
- *Thanks for installing TGX on my laptop. I am glad that I did so as I am presently on vacation in India. And TGX has outperformed any other remote work option for me by a margin (Hydra and EOD). The screen update is almost real time.* – Houston, TX based geophysicist
- *I love TGX, it has been a life saver for me and it is very stable. I have been on the same session and I just logout and back in and my work area is still there.* – Houston, TX based geophysicist

Questions?