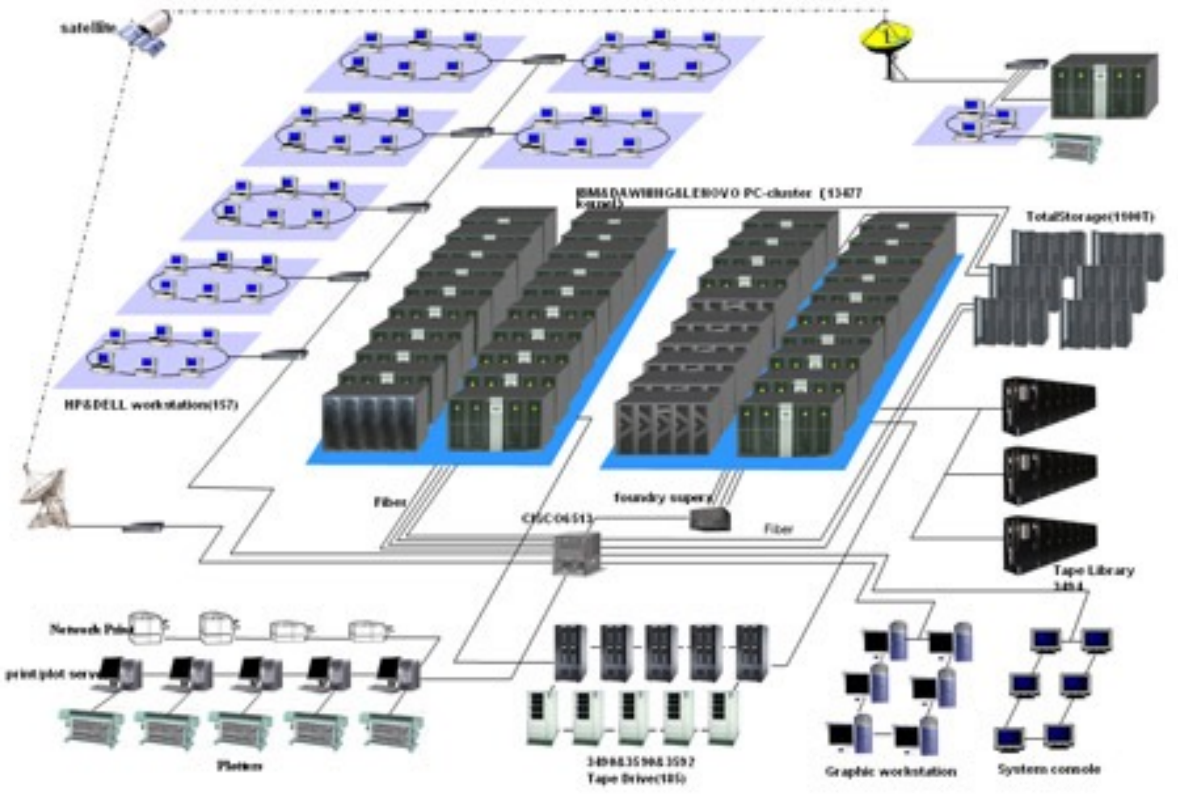
JavaSeis Cloud: Cloud-Native Seismic I/O

Outline

- The energy industry knows big data
- Parallel synchronous computing with MPI on large HPC systems
- ”Lift and Shift” to the Cloud is a reasonable first step
- Shift to Synchronous-MPI PLUS Asynchronous Microservices
- JavaSeis Cloud: Cloud-native seismic I/O
- Example: Distributed Acoustic Sensing processing
- Data science integration is easy in the Cloud
- It works !

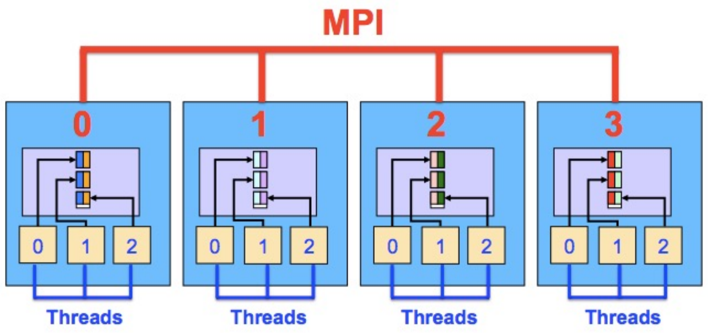
The energy industry knows big data

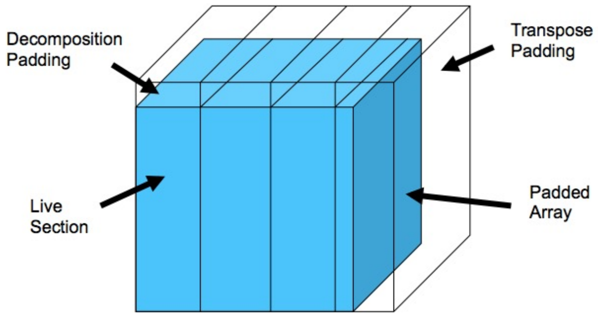
- Seismic data processing and imaging has driven the HPC market for decades
- Petabyte scale I/O and computation
- Our algorithms, workflows, and HPC resources are optimized for our business
- We excel at building and using systems for high performance networking, I/O, and computation
- The dominant computing paradigm is Distributed Synchronous Parallel computing (e.g., MPI)

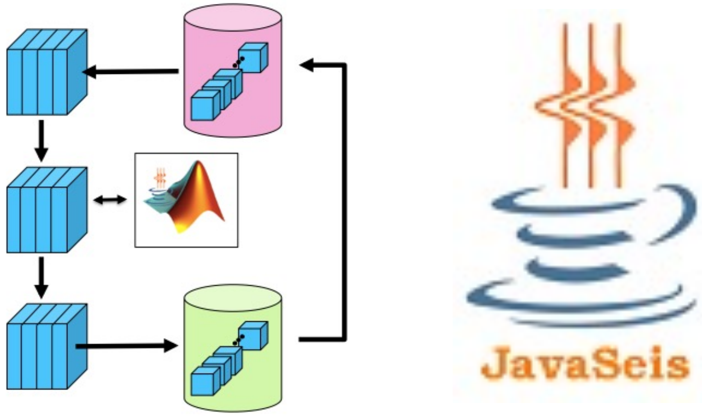


Example: JavaSeis Parallel Distributed Arrays

- Public domain (JavaSeis.org)
- Hosted on SourceForge and GIThub
- Object oriented SIMD synchronous parallel design
- High level abstractions for geophysics
- Parallel input, Parallel computation, Parallel output
- Scalable to millions of tasks
- 20 years of application development
- We will keep this !

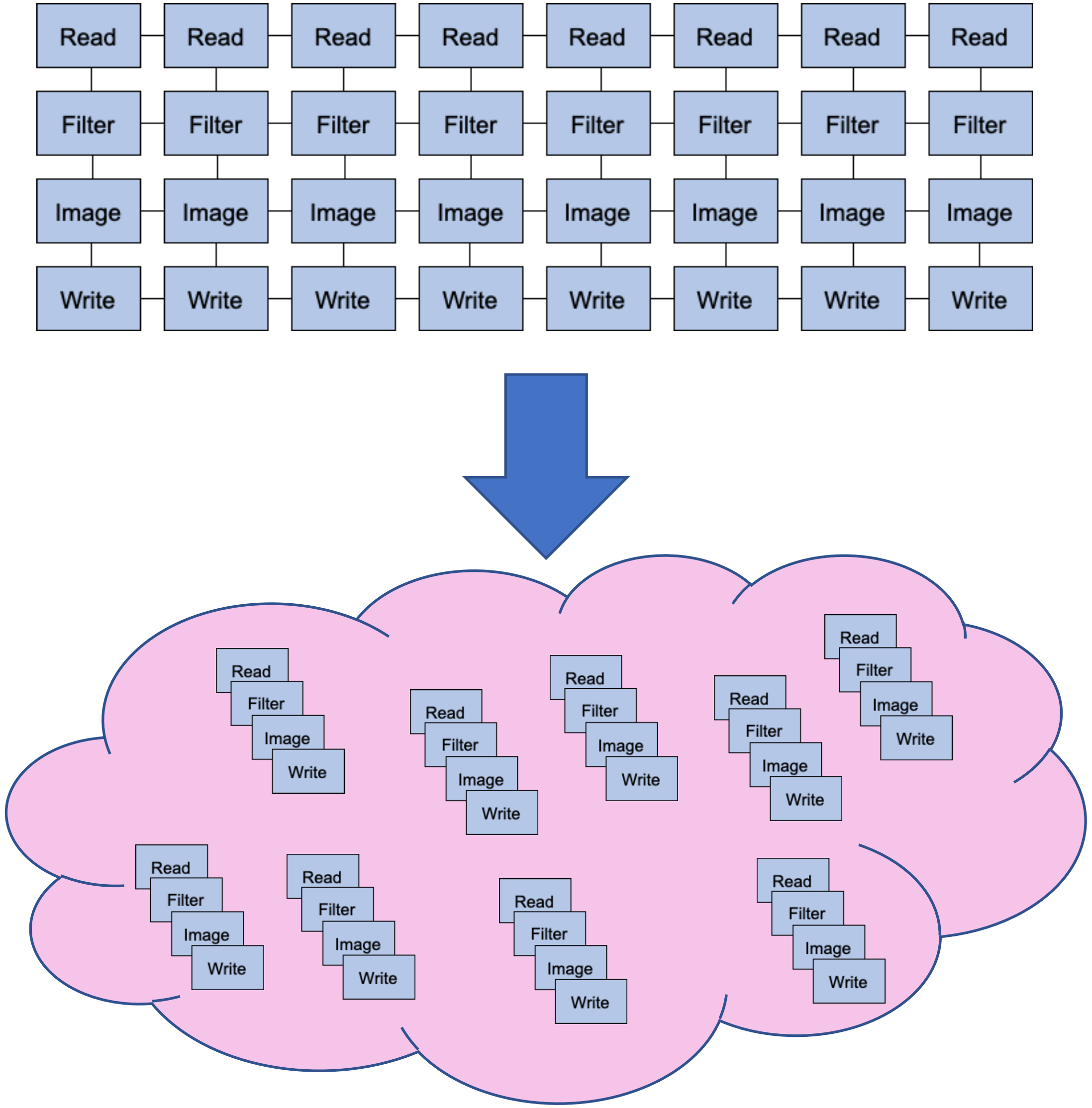






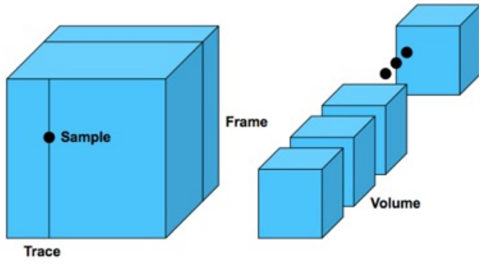
Asynchronous Task Parallel Computing

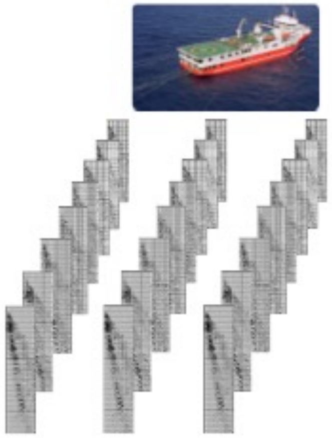
- Augment synchronous MPI style with a model for "stateless" independent tasks
- Supplement high-performance parallel file systems with distributed key-object storage
- Provide infrastructure for starting, monitoring, and completing a "cloud" of tasks
- But we need key-object storage and task-based I/O !

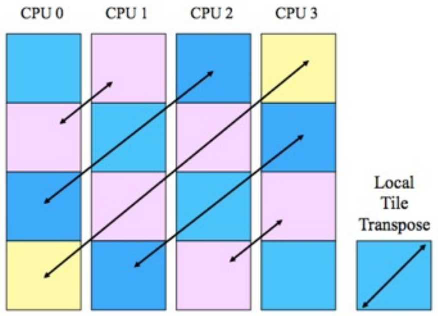


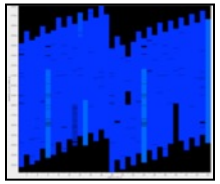
JavaSeis Parallel I/O

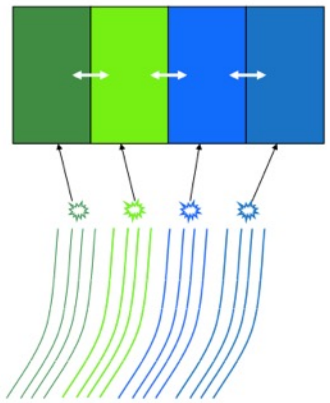
- JavaSeis I/O provides persistent storage of multi-dimensional arrays
- Sparse array, sparse disk, and compression algorithms
- Hierarchical tiled transpose algorithms for sorting
- Can utilize Terabytes of node memory for caching
- Multi-dimensional sorting at streaming I/O speeds
- Makes data management simple from user perspective
- One input dataset, one output dataset, one job, 1000s of nodes

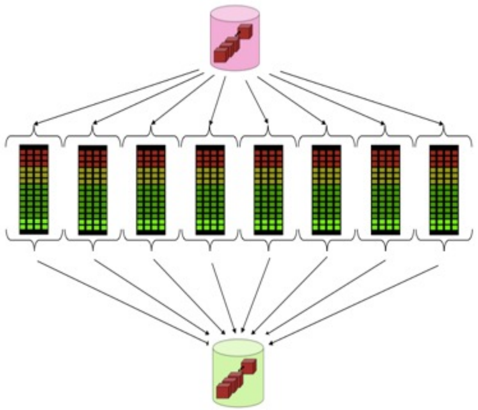






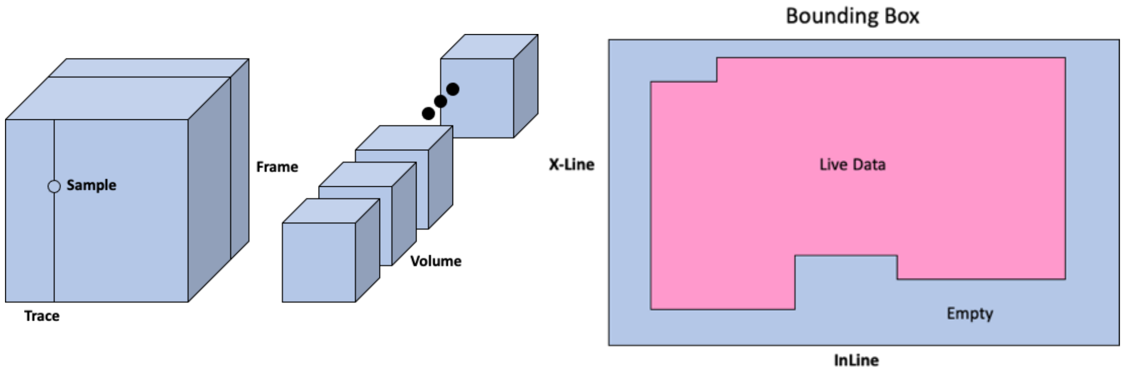


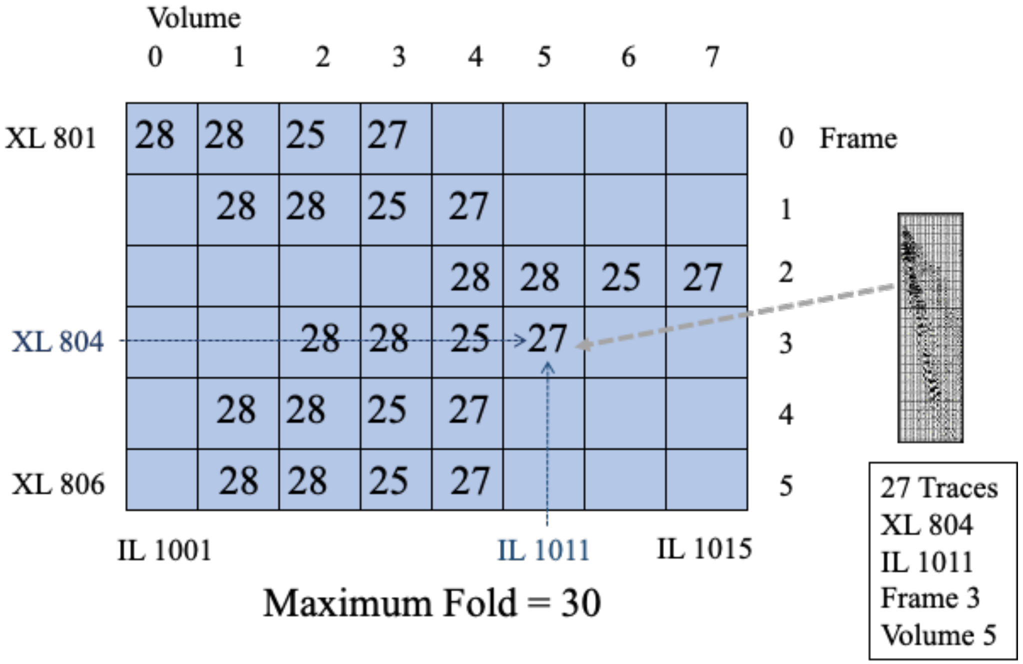




JavaSeis Model for Pre-Stack Seismic Data

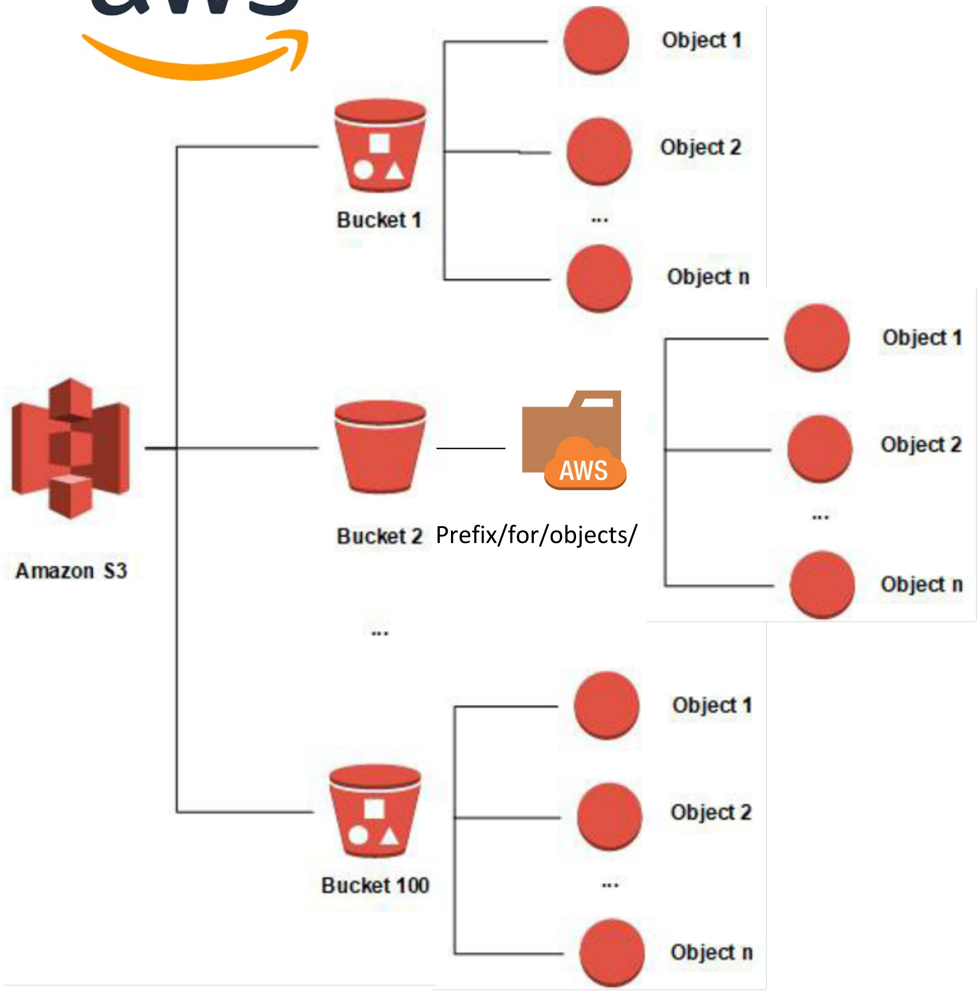
- Represent multi-dimensional seismic data as a collection of 3D volumes
- A bounding box defines the limits of the dataset
- A “Fold Map” records the number of traces in each cell of the bounding box
- JavaSeis provides an interface that supports this model
- Build a key-object interface





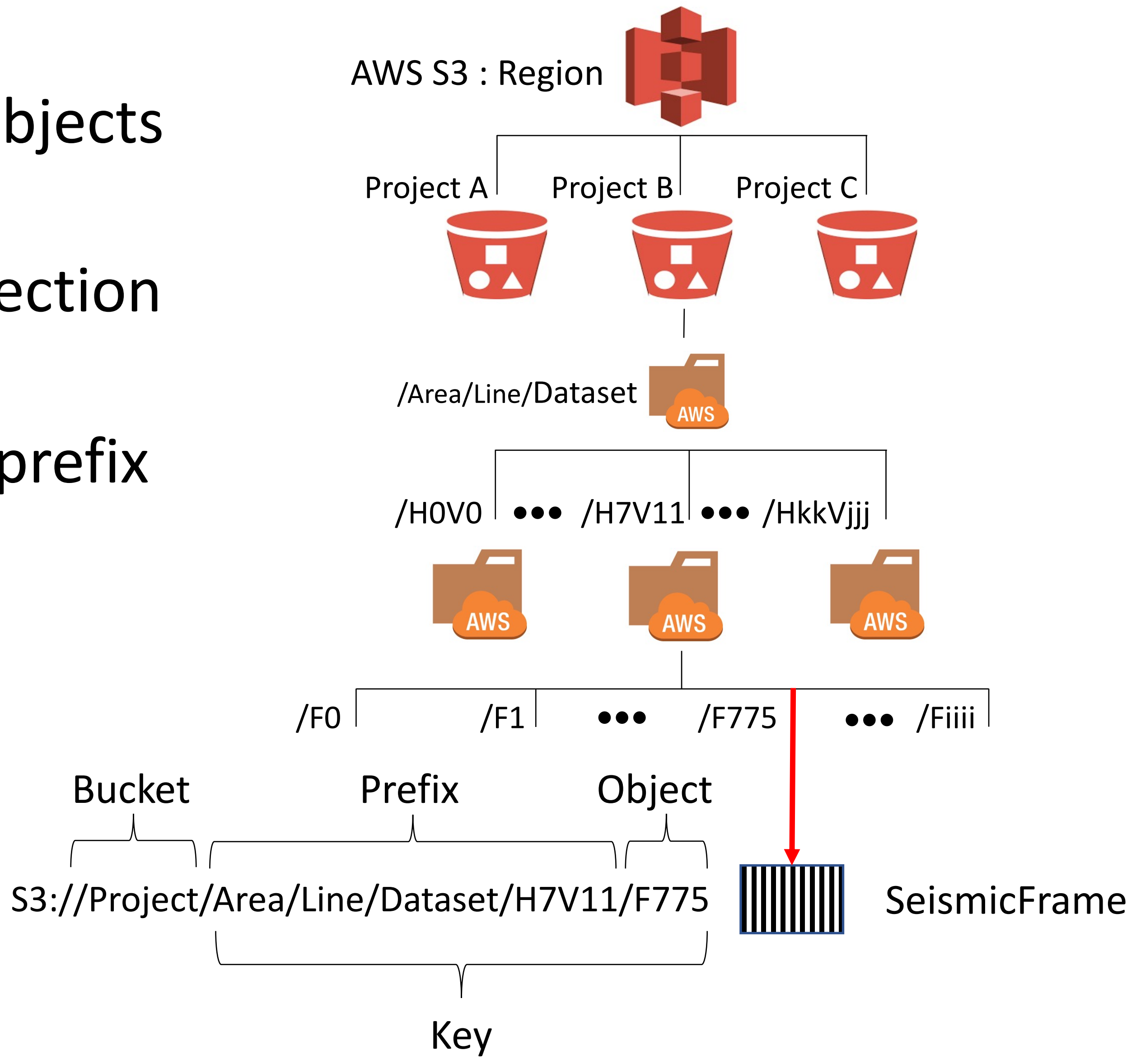
Key-Object Storage – AWS Simple Storage Service

- Optimized for search and horizontal scalability
- Built on "Map-Reduce" style file systems
- Access by well known public API's
- In AWS S3 a "Bucket" represents the access point
- Folders represented by a "Prefix"
- Key for object is Bucket+Prefix+Name: s3://Bucket/Prefix/ObjectName
- Horizontal scaling achieved across buckets and prefixes



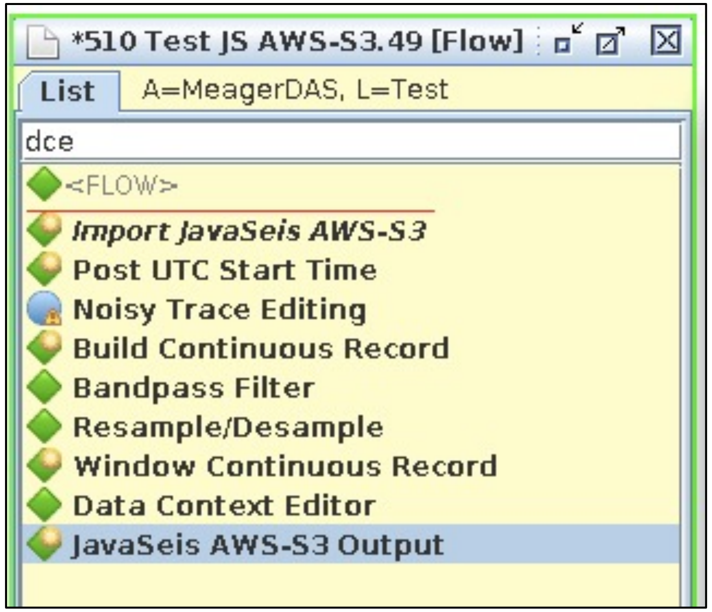
JavaSeis Cloud Interface for AWS S3 (JSC-S3)

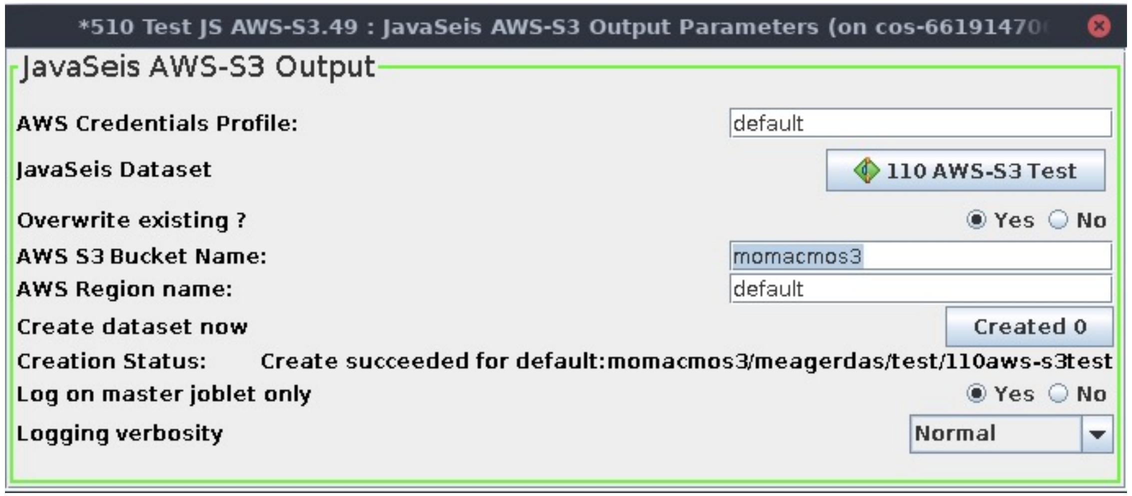
- JSC-S3 Interface uses the same objects as JavaSeis
- N-D data is represented as a collection of volumes
- A JSC-S3 dataset is mapped to a prefix in a bucket
- Hypercube/Volume mapping: DataSet/HhhVvvv
- Frames are mapped to: DataSet/HhhVvvv/Fffff



JavaSeis Cloud for S3 Performance

- It’s Java - Once we have an interface implementation we can use it !
- 1-week to implement and test
- Initially deployed in SeisSpace™
- Achieves ~10 MB/s read/write per thread
- 32 jobs in SeisSpace Essentials scales to 320 MB/s
- Lambda “Free Tier” with 1000 concurrent tasks scales to 10 GB/s
- $25 per TB-Read-Write





MoMacMo Limited

Geophysics for the common good

OUR PROJECTS

Technology Development

Geophysical technology development for hazard identific identification and infrastructure monitoring

Project planning and operations

Plan and operate projects that use geophysical technology to improve our community and world

Collaboration and Education

Collaborate with other non-profits, for-profits, and universities on projects that support the MoMacMo mission

Geologic Hazard Monitoring and Prediction

- Earthquakes
- Volcanic eruptions
- Landslides
- Glacial outburst
- Subsidence
- Pipeline leaks

The Mt. Meager DAS Experiment

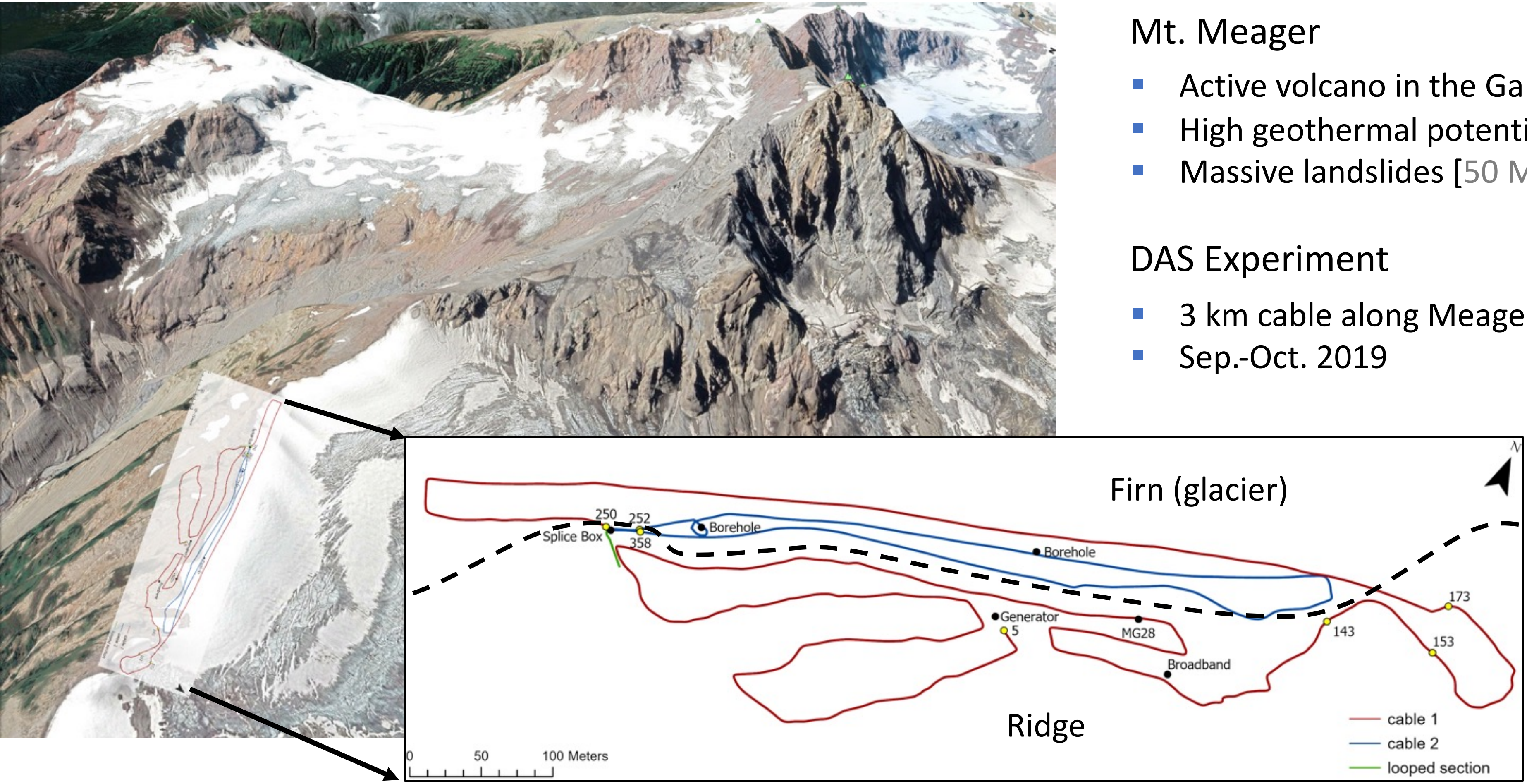
Sara Klaasen ETH, Andreas Fichtner ETH, Jan Dettmer U. Calgary

Mt. Meager

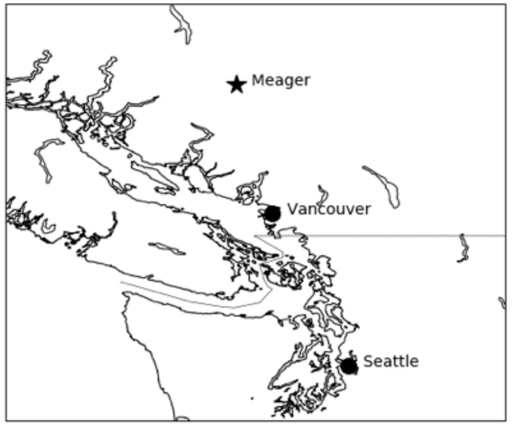
- Active volcano in the Garibaldi range
- High geothermal potential
- Massive landslides [50 Mio m^3 in 2010]

DAS Experiment

- 3 km cable along Meager ridge at 2100 a.s.l.
- Sep.-Oct. 2019



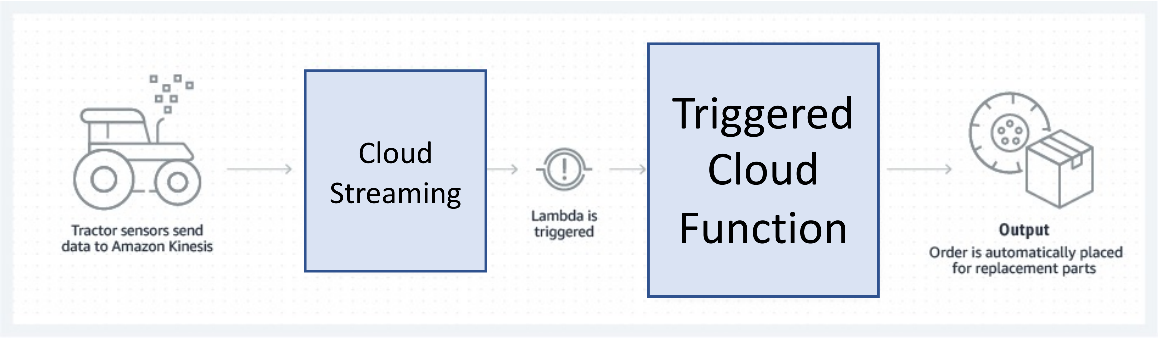
Cable layout on Meager ridge



Location of Mt. Meager

MoMacMo DAS Processing

- DAS systems can generate Terabytes of data per day
- Processing and management of this data is a daunting task
- MoMacMo has technology to create multi-scale versions of the data at a range of scales - days to milliseconds
- We use cloud storage (AWS and Azure), Apache Spark Machine Learning, and JavaSeis.org technology for parallel I/O and compute
- Our systems can generate interpretive datasets from raw DAS data at rates at very low cost

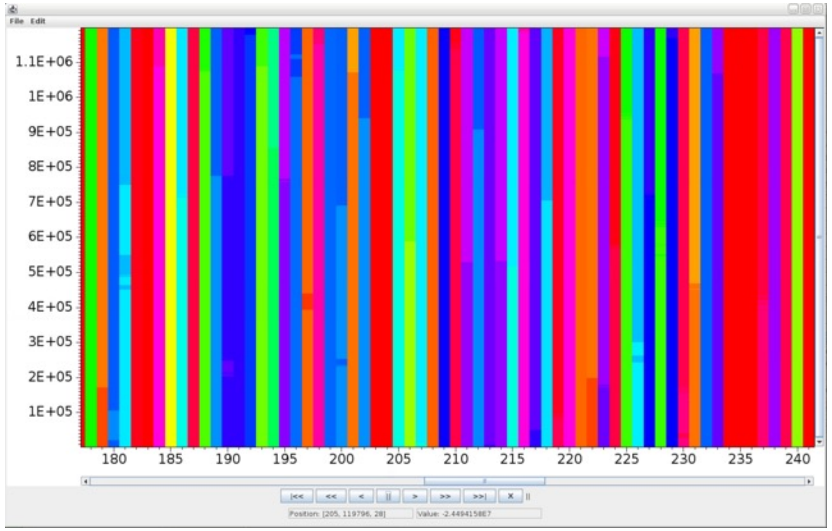


- DAS HDF5 files are loaded to cloud storage
- Event triggers a cloud function to read the HDF5 and convert to JavaSeis format
- Metadata from HDF5 is used to establish the time segment of the data
- The JavaSeis dataset “grows” in real time and can be viewed and processed any time
- QC tools allow us to watch and monitor data and performance from any device
- Cloud Machine Learning for event detection and location

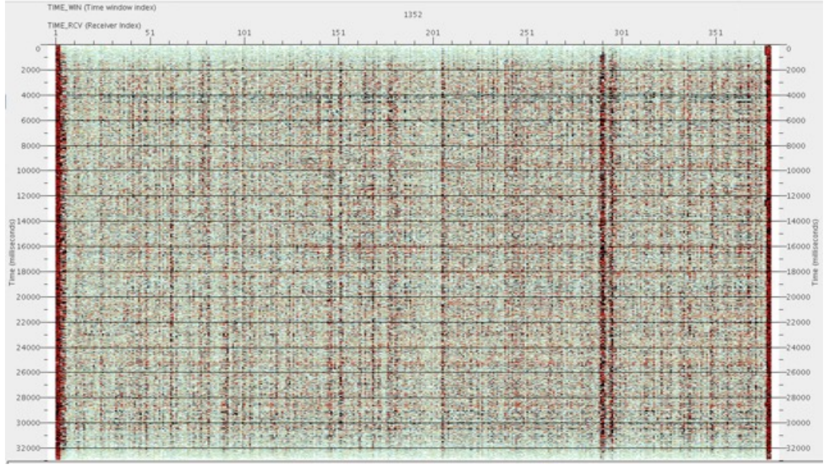
Mt Meager DAS Data Processing

- Data were loaded from University of Calgary servers in Canada
- 30 days of 1 msec recording on a 3 km fiber
- We used the Linux AWS Command Line Interface:
  `aws s3 cp *.h5 s3://glacier-das`
- 2500 files, 0.8 GB per file, 1.8 TB total, transferred in 8 hours
- Raw data is converted to scaled differential phase in windowed time
- Total conversion and transfer elapsed time was 12 hours

Raw integer data

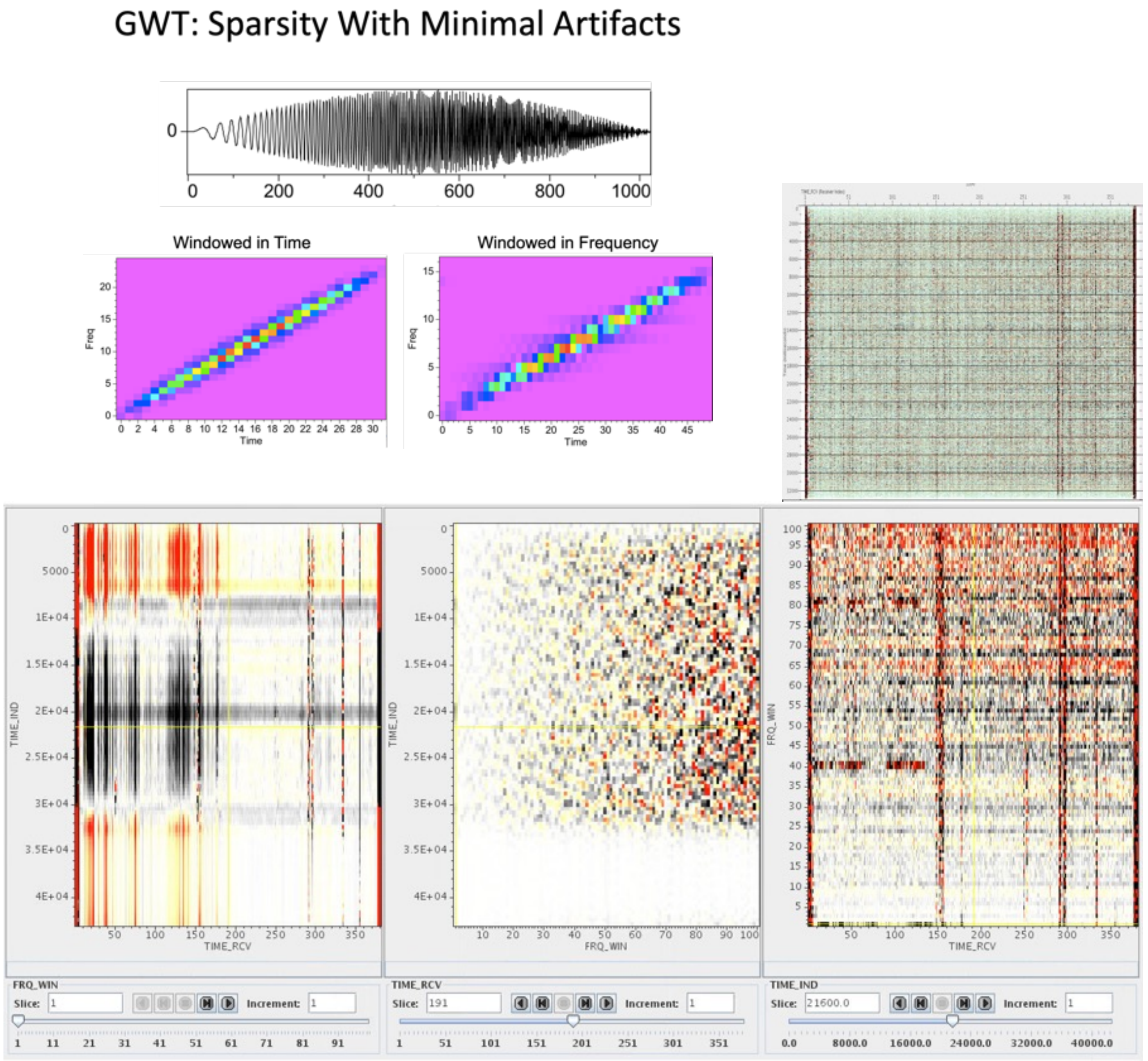


Scaled differential phase

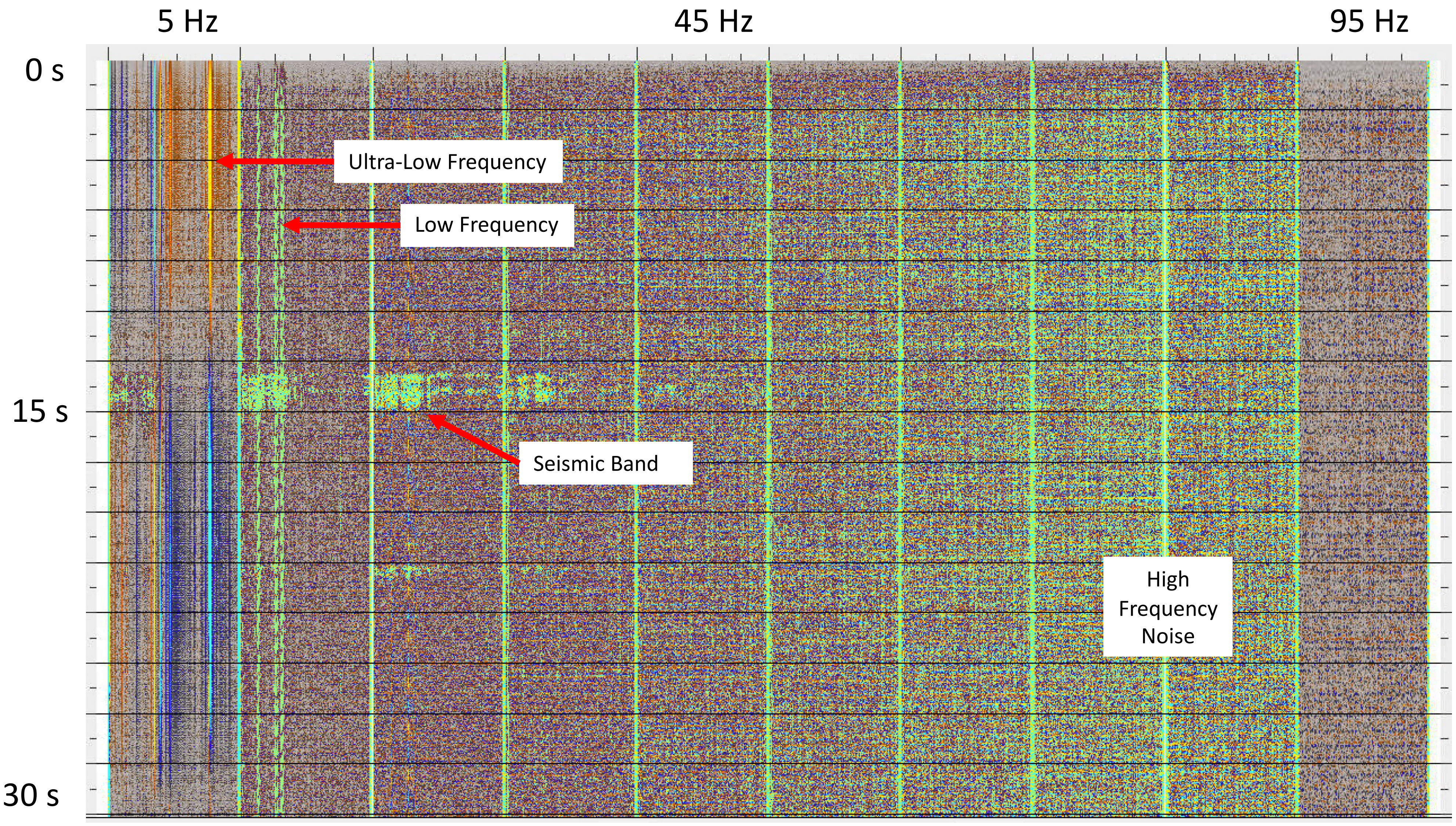


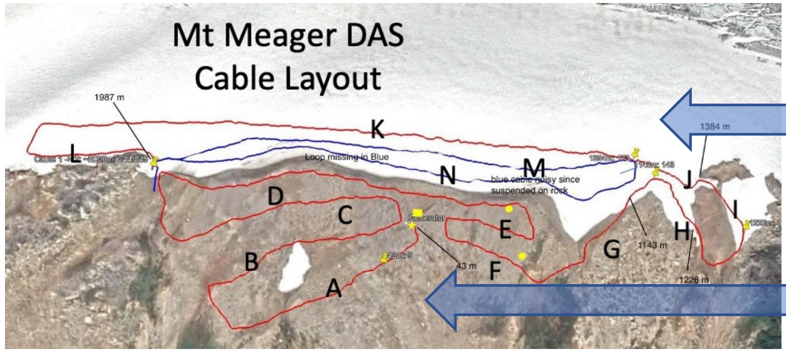
Multi-Scale Decomposition for DAS Data

- Based on the Generalized Windowed Transform (GWT, Mosher 2011)
- Framework for wavelet style transforms that avoid sub-band aliasing
- Real to real transform in "Frequency Bands"
- Mathematical coupling between sub-bands retains ALL frequencies



Meager DAS Time-Frequency Decomposition



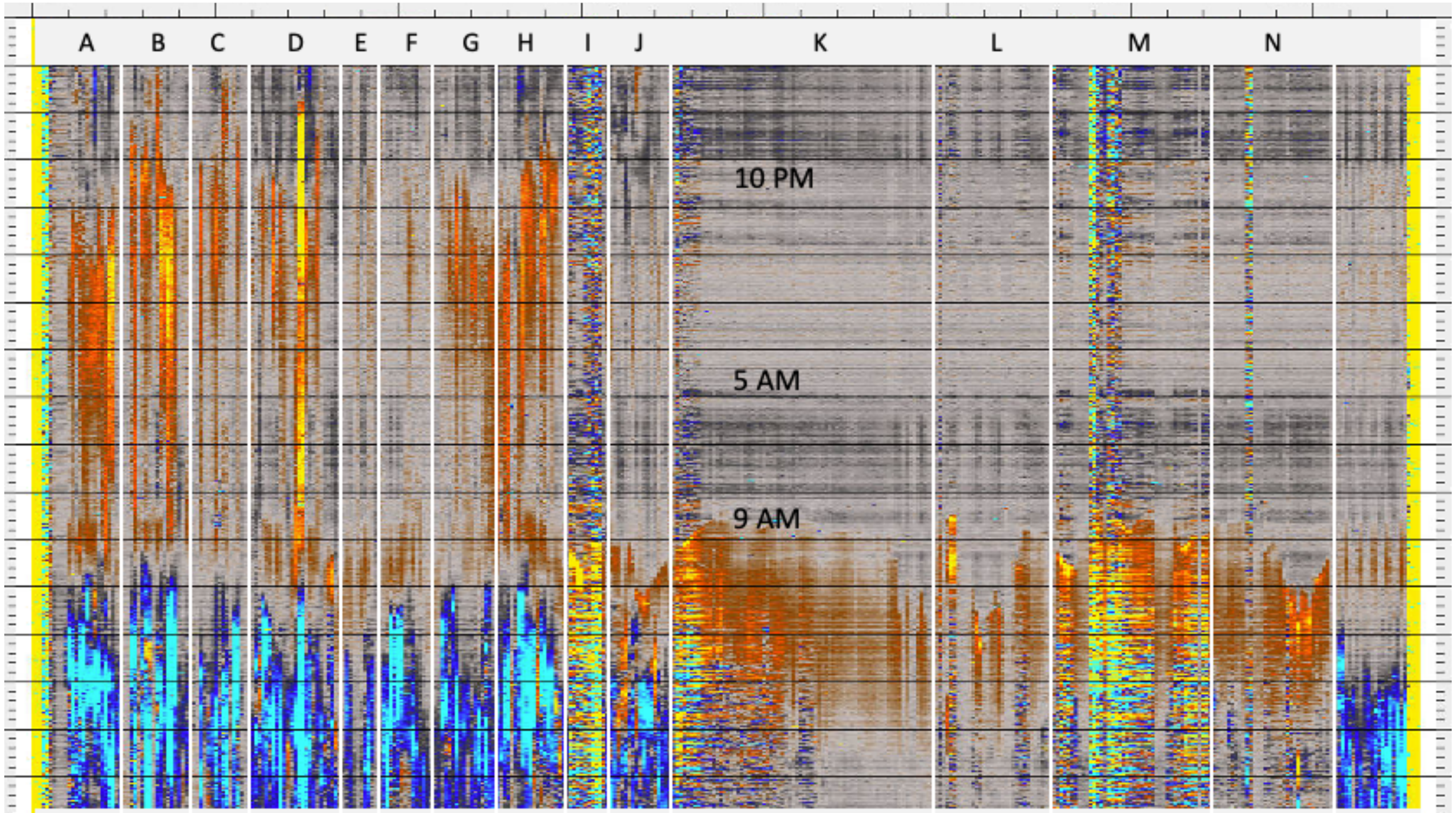


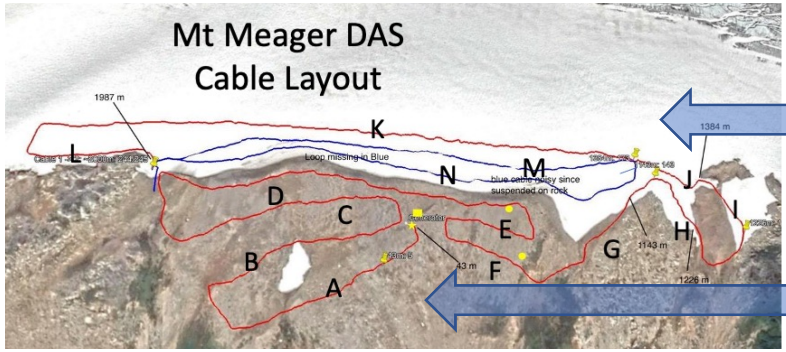
Glacier (K-N)

Ridge (A-J)

Low Frequency Strain

- Zero to 1 Hz
- Temp induced
- Warm = expansion
- Cold = contraction
- One UTC day
- Sunrise on glacier
- Wind on ridge



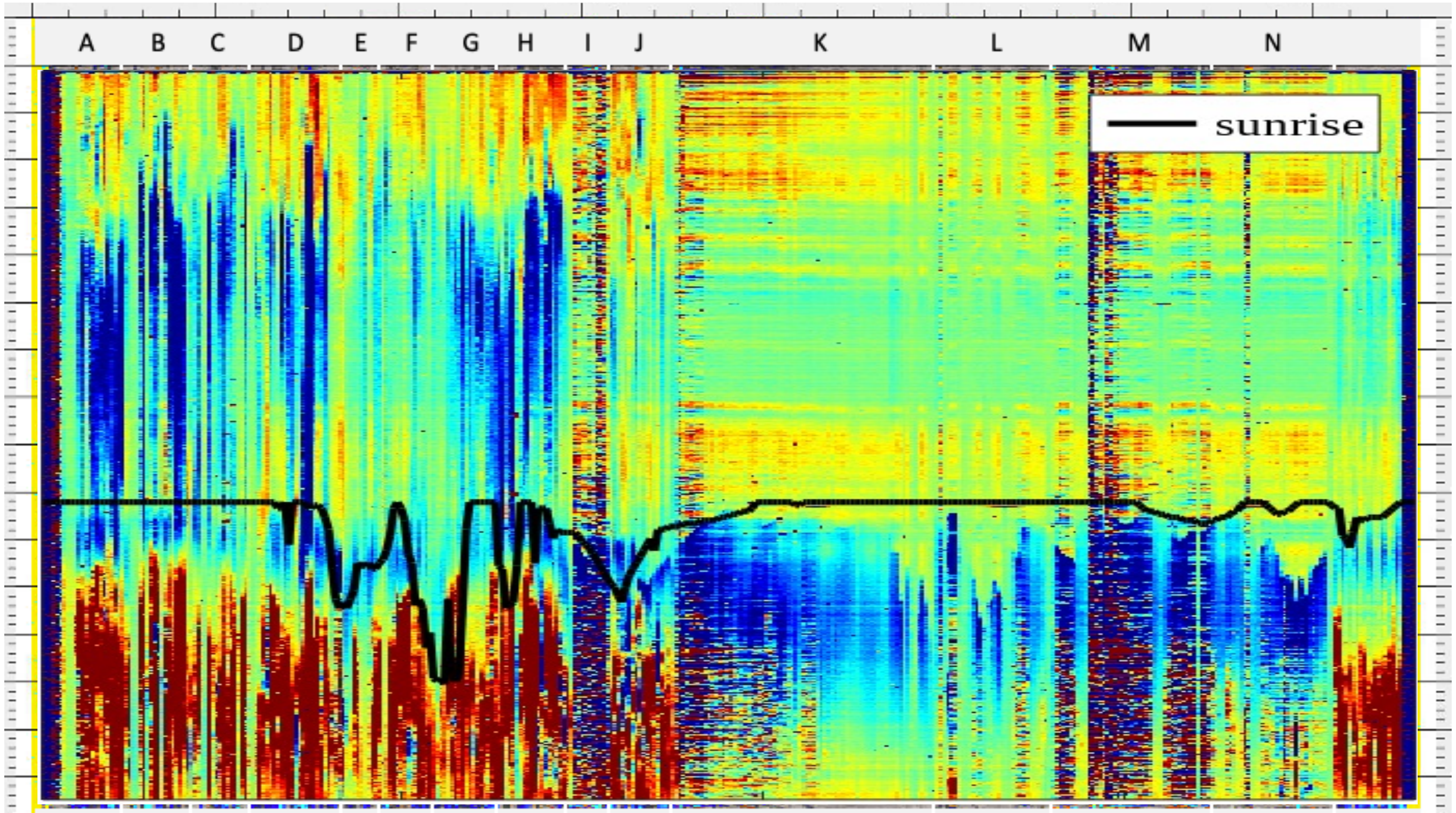


Glacier (K-N)

Ridge (A-J)

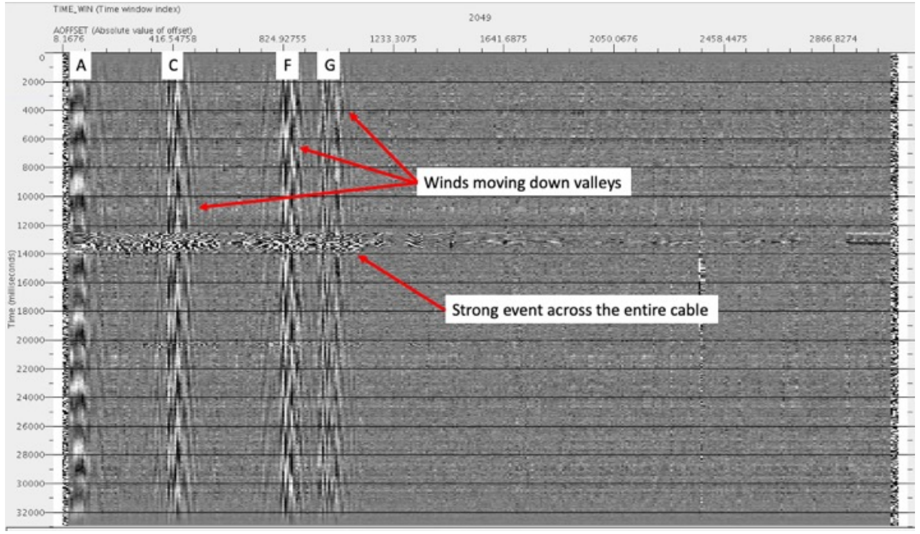
Low Frequency Strain

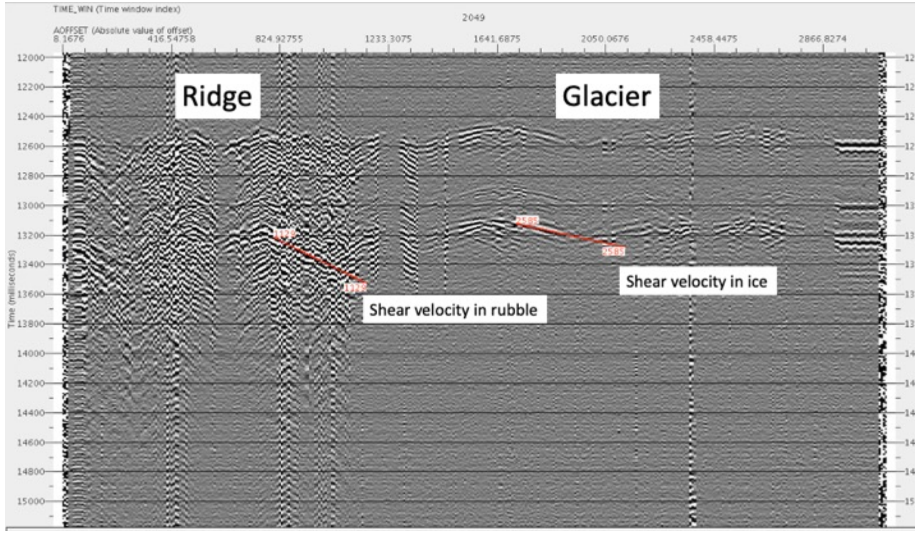
- Zero to 1 Hz
- Temp induced
- Warm = expansion
- Cold = contraction
- One UTC day
- Sunrise on glacier
- Wind on ridge



Numerous Seismic Events in the 5-50 Hz Band

- We see low frequency wave trains in valleys with 1-5 m/s velocity
- Events that occur on all channels
- Complex waveforms traveling at near surface shear velocities
- Events appear to originate on the glacier and on the slopes
- Waveform and source mechanism analysis are underway

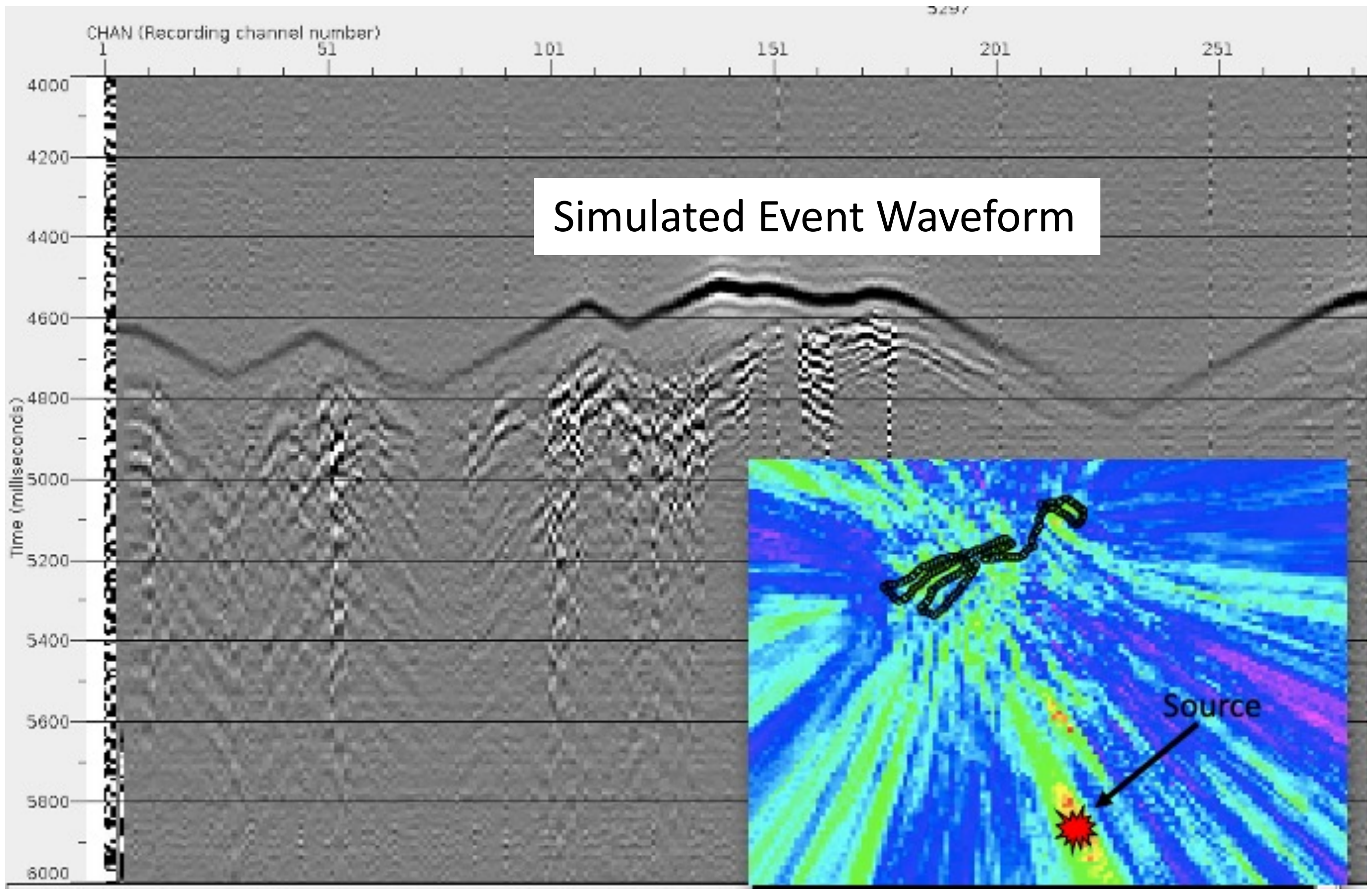




Event Location with Distributed Acoustic Sensing

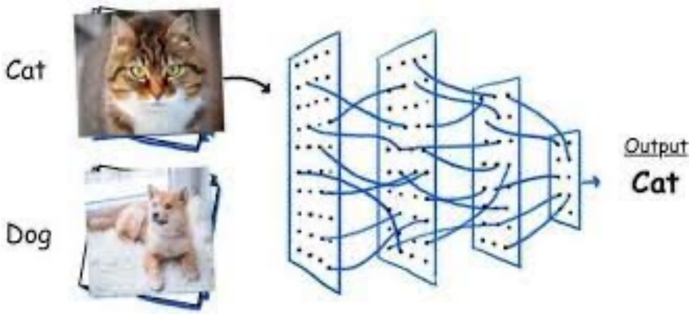
- DAS detects broad-band differential strain on fibers
- Events generate energetic surface waves in the 5-50 Hz range
- We use interferometry to provide event energy density maps
- Detection and location is time consuming
- Could we use Machine or Deep Learning ?

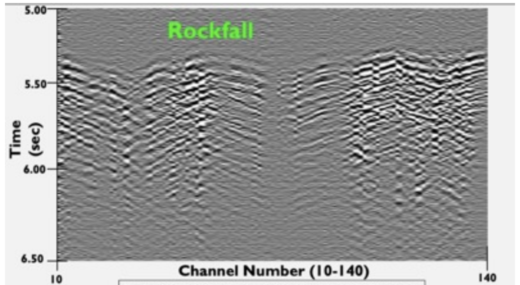
2D Kirchhoff Interferometry



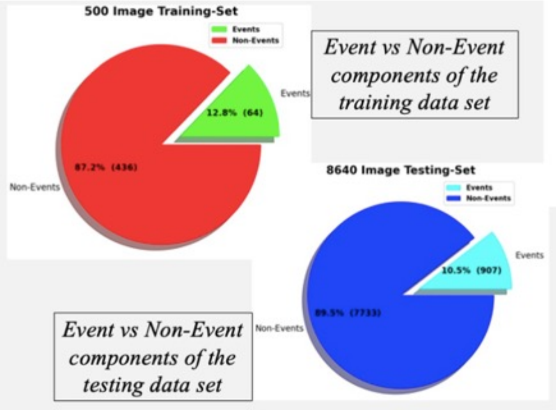
Deep Learning Example: DAS Event Detection

- The Cat vs Dog Tutorial
- VGG16 Deep Learning Network with Transfer Learning
- Train using events labeled by experts
- Predicted 900 events using a full day of recording
- Matches performance of human experts
- Work done by undergrad for senior thesis



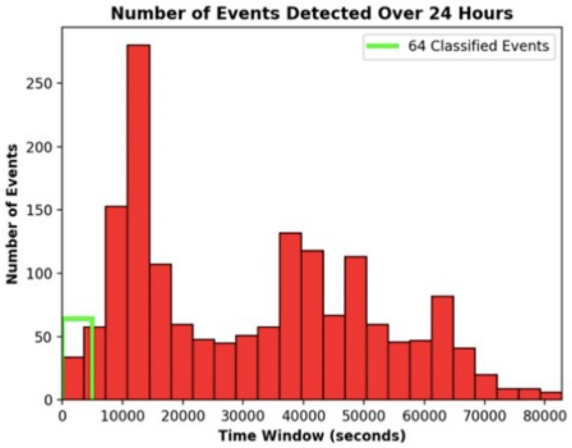


Rockfall captured on channels 10-140. Used for training the predictive model



Event vs Non-Event components of the training data set

Event vs Non-Event components of the testing data set



Number of energy events detected by the neural network from 8640 frames over 24 hours on day 11

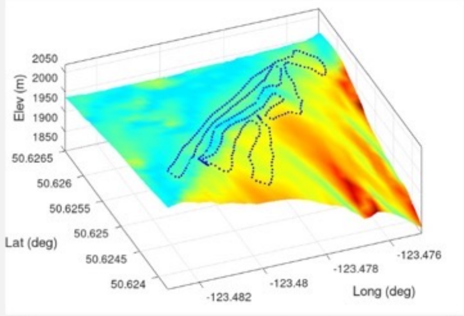
MACHINE LEARNING AND DISTRIBUTED ACOUSTIC SENSING DATA FOR RISK ASSESSMENT: GRAVITY DRIVEN GEOTECHNICAL HAZARDS

Jacob Mish[1]

[1]Department of Geophysics, Research Mentors: Dr. Robert. J. Ferguson

Introduction

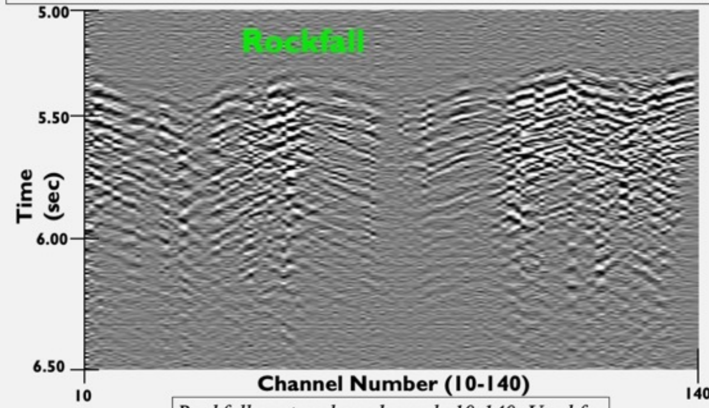
- Detect real-time rockfalls using machine learning
- Distributed Acoustic Sensing (DAS) data from fiber optic cables on Mt. Meager
- Monitor rockfalls near important workers and assets
  - Geothermal
  - Hydroelectric
  - Transport Infrastructure



DAS station map on Mt. Meager

Methodology

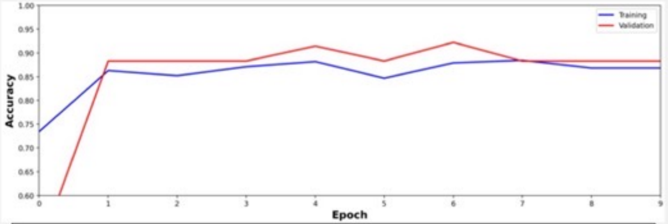
- Isolate microseismic band with the Generalized Window Transform: 10-40 Hz
- Train a convolutional neural network to classify images of energy events & non-events



Rockfall captured on channels 10-140. Used for training the predictive model

Conclusions

- Transfer learning: Fast & Effective Training
- Lines and Features: Easily Extracted and Identified
- Highly Successful: 80% accuracy for first model
- Real-time: Instantaneous classification

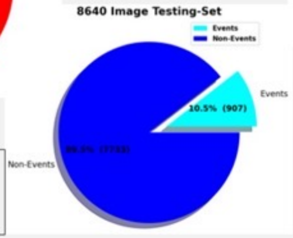


Classification accuracy of the machine learning model over 10 epochs

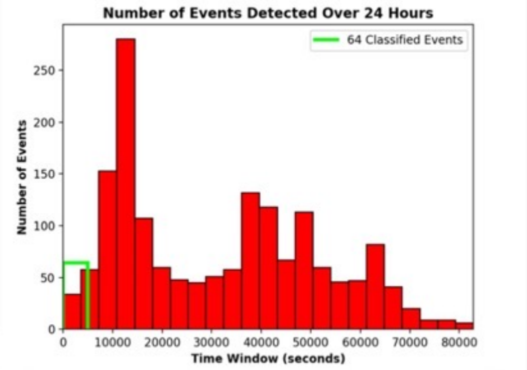
Examples

- Training Dataset: 64 Rockfalls in hour 1.5
- Testing Dataset: 907 Rockfalls in 24 hours

Event vs Non-Event components of the training data set



Event vs Non-Event components of the testing data set



Number of energy events detected by the neural network from 8640 frames over 24 hours on day 11

Acknowledgements

- Jan Dettmer – Department of Geoscience, University of Calgary
- Chuck Mosher - MoMacMo LTD.
- Glyn William-Jones – Department of Earth Sciences, Simon Fraser University

UNIVERSITY OF CALGARY

HALLIBURTON
Landmark

SEISMIC ENGINE

JavaSeis

Mt Meager DAS Experiment Learnings

- Multi-scale decomposition of DAS data provides a wealth of domains to monitor physical processes
- The low frequency DAS response has intriguing correlation with weather and temperature
- Numerous seismic events in the 5-50 Hz band
- Deep Learning networks can efficiently detect seismic events in large amounts of data
- Interferometric imaging allows us to infer event locations
- These components could be used to create a real time hazard detection system using DAS and cloud platforms

JavaSeis Cloud: Conclusions

- The energy industry knows how to process and use big data
- Our traditional HPC applications can be shifted to the cloud and continue to provide value
- Shifting appropriate work loads to a cloud native framework has potential to reduce both manpower and project costs
- We have implemented the JavaSeis pre-stack data interface for Amazon S3 and Azure Blob Storage
- Performance matches high end parallel file systems at significantly lower costs
- Bringing seismic data closer to data science will transform seismic processing