Big Data Framework for Synchrophasor Data Analysis

Pavel Etingov, Jason Hou, Huiying Ren, Heng Wang,
Troy Zuroske, and Dimitri Zarzhitsky

Pacific Northwest National Laboratory

North American Synchrophasor Initiative Group Meeting, April 24-26, 2018, Albuquerque, NM

Project team

- Project is supported by the DOE through the GMLC program

- PNNL
  - Pavel Etingov
  - Jason Hou
  - Huiying Ren
  - Heng Wang
  - Troy Zuroske
  - Dimitri Zarzhitsky

- Partners
  - LANL
  - LBNL
  - BPA

Project goals

- Develop a framework for PMU big data analysis
  - Event detection
  - Abnormalities detection
  - Improved situational awareness
  - System identification (learning system dynamic behavior)
  - Advanced visualization
- Framework is based on the cloud technology and distributed computing:
  - PNNL institutional cloud system or Microsoft Azure
  - Apache SPARK for distributed big data analysis and Machine Learning (ML)

PNNL cloud infrastructure

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

- PNNL cloud is based on OpenStack (a free and open-source software platform for cloud computing)
- Cloudera Apache Hadoop Distribution:
  - Apache Spark (an open source cluster computing framework)
  - Apache Hive (a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis)
  - HBase (an open source, non-relational, distributed database)

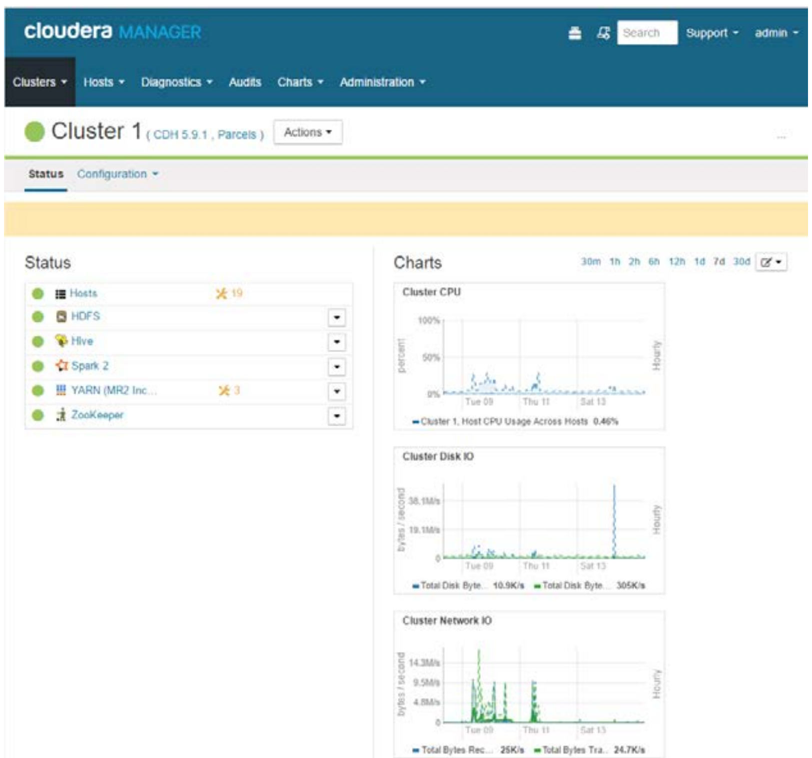
Apache Spark

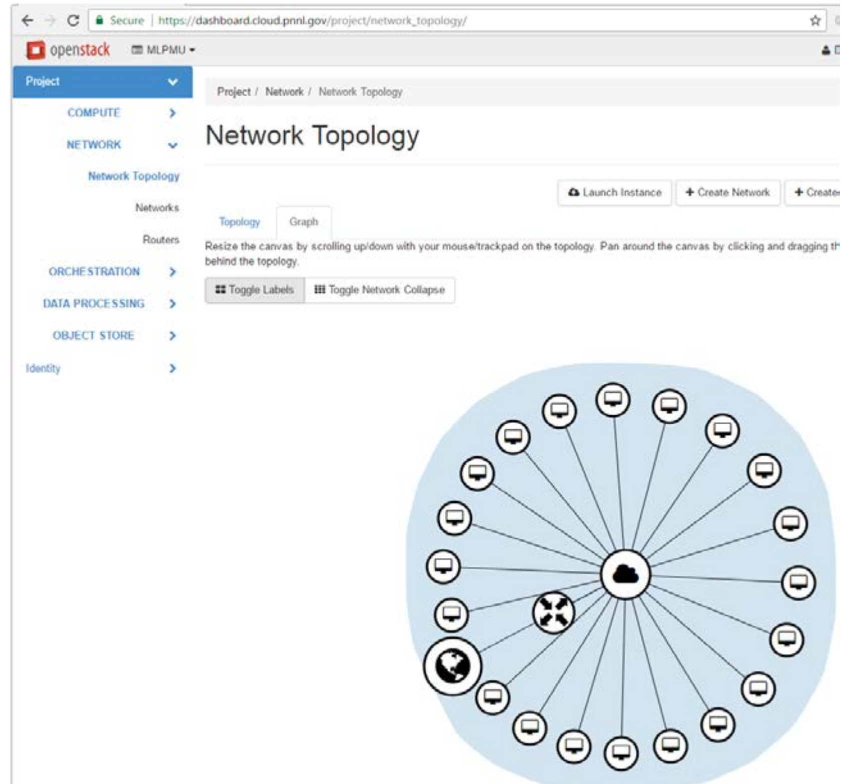
- Large scale parallel data processing framework
- Extremely powerful (up to 100x faster than Hadoop)
- Large datasets distributed across multiple nodes within a computer cluster
- Support real time data stream
- Built-in Machine Learning library
- Support different languages (Scala, Java, Python, R)
- Support different data sources (SQL, Hive, HBase, Cassandra, Oracle, etc)
- Open source and free
- Available through public cloud services (Amazon AWS, Microsoft Azure, IBM, etc) and through new PNNL institutional cloud system.

Spark research cluster based on PNNL cloud

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

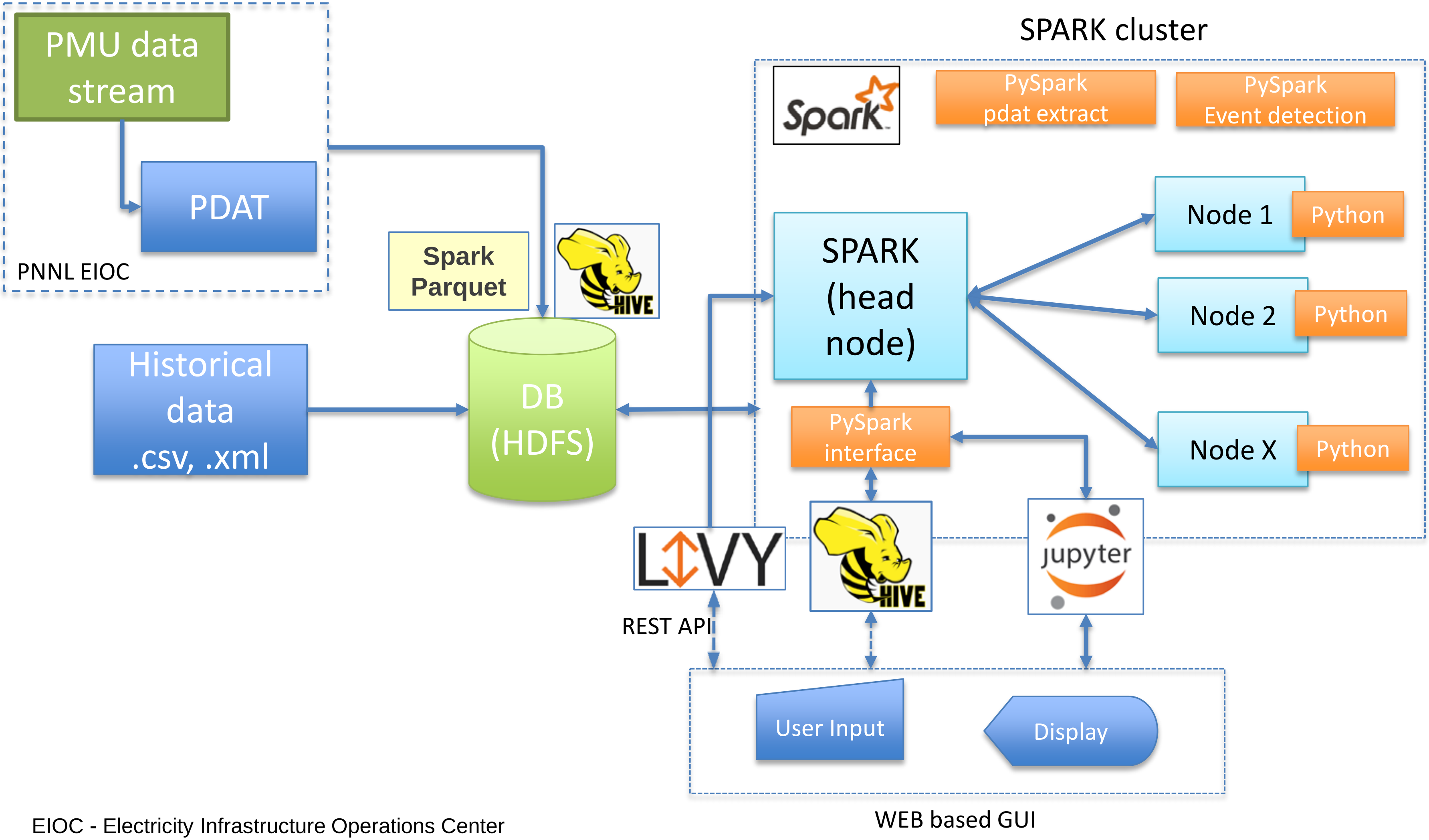
- Current configuration
  - 20 nodes
  - RAM 512 Gb
- Recently upgraded to Spark 2.2





- Cluster will be upgraded to 1 Tb RAM

Cloud based ML-PMU Framework



EIOC - Electricity Infrastructure Operations Center
HDFS- Hadoop Distributed File System

PMU data stream

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

- PNNL receives PMU data stream from Bonneville Power Administration
  - 12 PMUs
  - Multiple channels (Voltage and Current Phasors, Frequency, ROCOF)
- PMU Data stored in PDAT format
  - PDAT format developed by BPA
  - Based on IEEE Std. C37.118.2-2011
  - Binary files
  - Each file contains 1 minute of data
  - One file ~ 5 MB

Data frame organization defined by IEEE C37.118.2

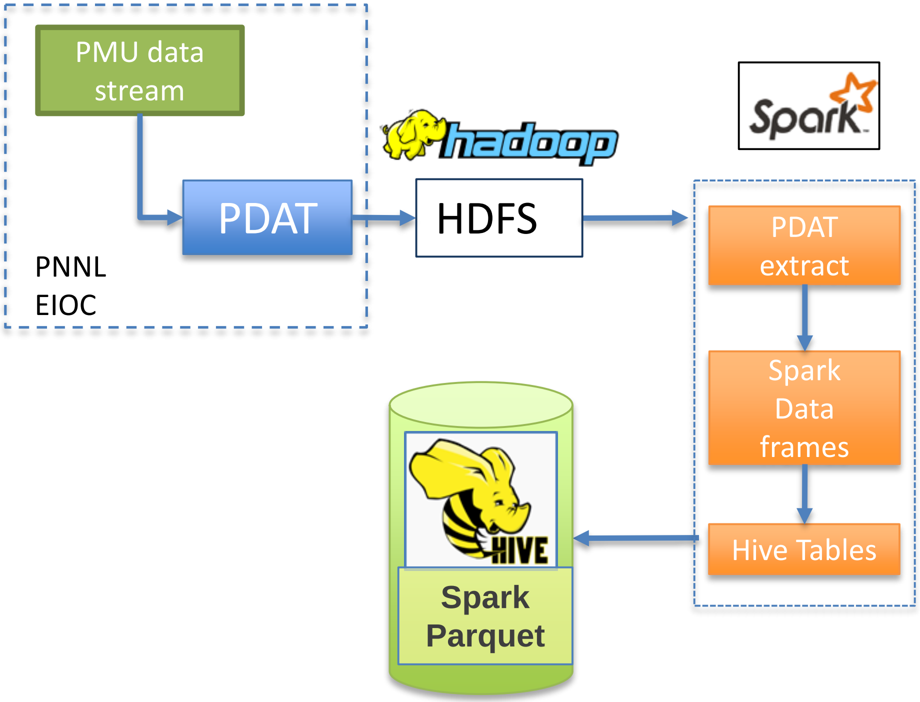
No.	Field	Size (bytes)	Comment
1	SYNC	2	Sync byte followed by frame type and version number.
2	FRAMESIZE	2	Number of bytes in frame, defined in 6.2.
3	IDCODE	2	Stream source ID number, 16-bit integer, defined in 6.2.
4	SOC	4	SOC time stamp, defined in 6.2, for all measurements in frame.
5	FRACSEC	4	Fraction of Second and Time Quality, defined in 6.2, for all measurements in frame.
6	STAT	2	Bit-mapped flags.
7	PHASORS	4 × PHNMR or 8 × PHNMR	Phasor estimates. May be single phase or 3-phase positive, negative, or zero sequence. Four or 8 bytes each depending on the fixed 16-bit or floating-point format used, as indicated by the FORMAT field in the configuration frame. The number of values is determined by the PHNMR field in configuration 1, 2, and 3 frames.
8	FREQ	2 / 4	Frequency (fixed or floating point).
9	DFREQ	2 / 4	ROCOF (fixed or floating point).
10	ANALOG	2 × ANNMR or 4 × ANNMR	Analog data, 2 or 4 bytes per value depending on fixed or floating-point format used, as indicated by the FORMAT field in configuration 1, 2, and 3 frames. The number of values is determined by the ANNMR field in configuration 1, 2, and 3 frames.
11	DIGITAL	2 × DGNMR	Digital data, usually representing 16 digital status points (channels). The number of values is determined by the DGNMR field in configuration 1, 2, and 3 frames.
	Repeat 6–11		Fields 6–11 are repeated for as many PMUs as in NUM_PMU field in configuration frame.
12+	CHK	2	CRC-CCITT

Ongoing work

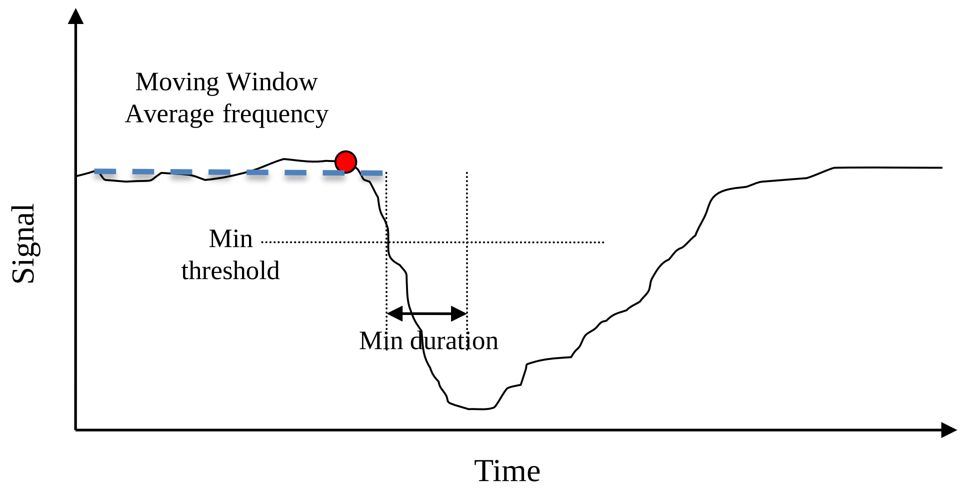
- Python (PySpark) modules:
  - PDAT data extraction
  - Data processing
    - Bad data
    - Missing points
    - Outliers
  - Event detection
    - Frequency events
    - Voltage events
  - Features extraction and analysis
    - Wavelet
    - Dynamic regression
    - Principal component analysis

PDAT data extraction

- Read information from PDAT and creates SPARK data frames
- Store information in Hive or Parquet tables
- Implemented in PySpark that allows parallel processing of multiple PDAT files
- Significantly increased performance
  - To read information for 1 hour takes about 20 seconds (20 nodes cluster)



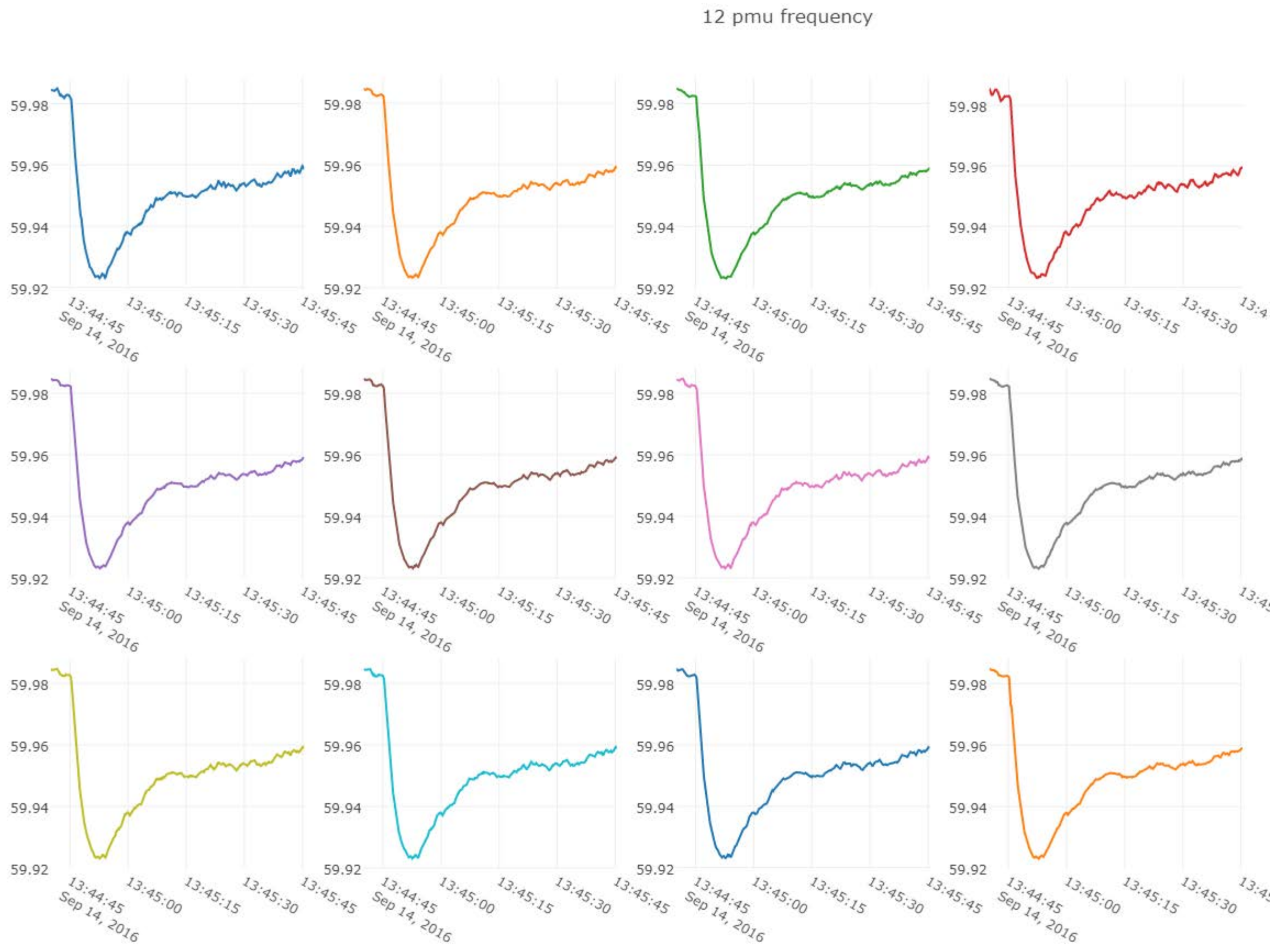
Event detection (threshold based)



- User specified
  - Delta frequency
  - Event duration
- Cross validation signal checks to avoid false alarms
- Spark usage significantly increases the computational throughput of the application
- Processing of 1 day takes about 5-7 minutes (processing the same dataset using a PC takes about 1 hour)

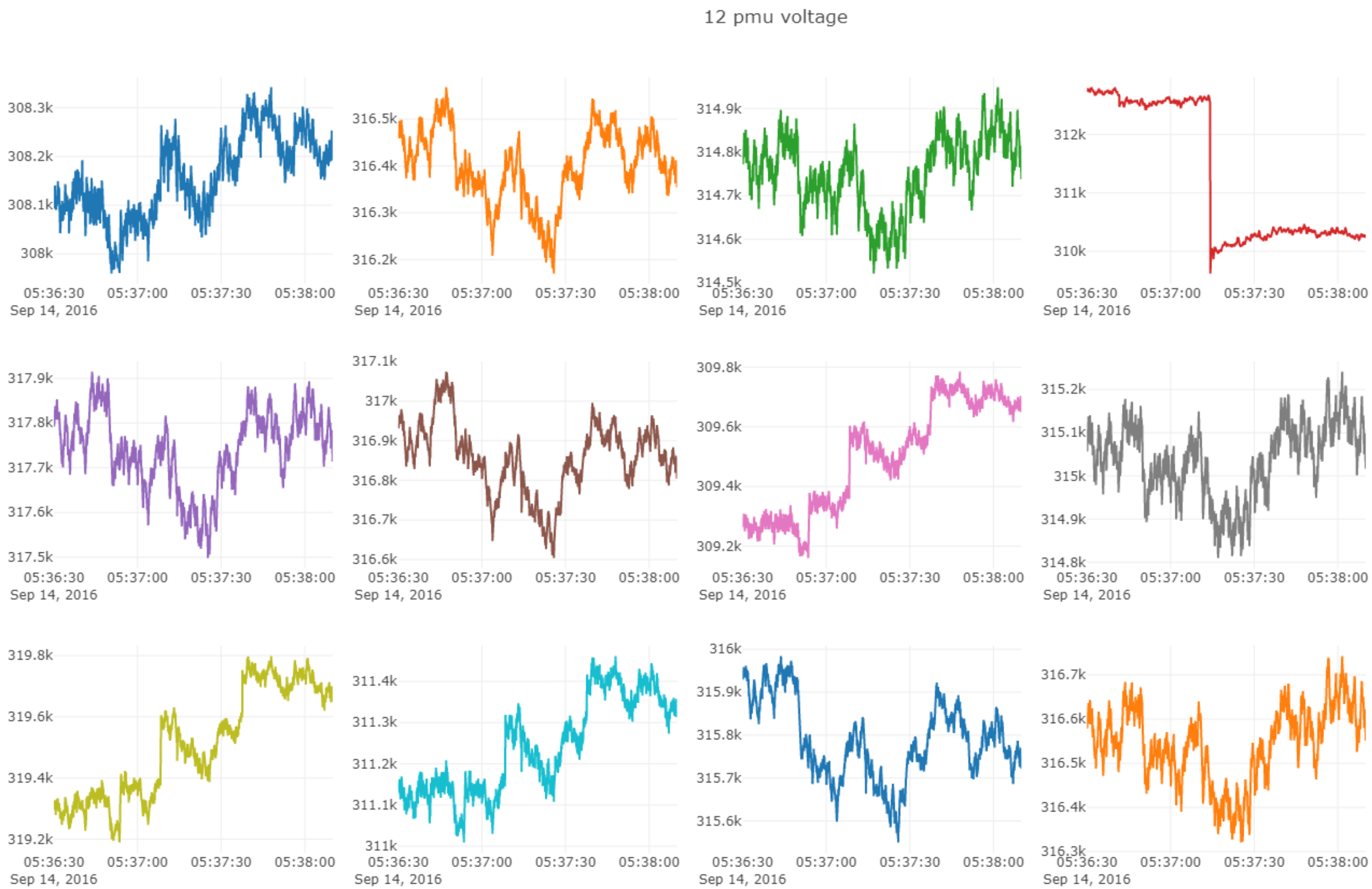
Examples of Detected events

- Frequency events



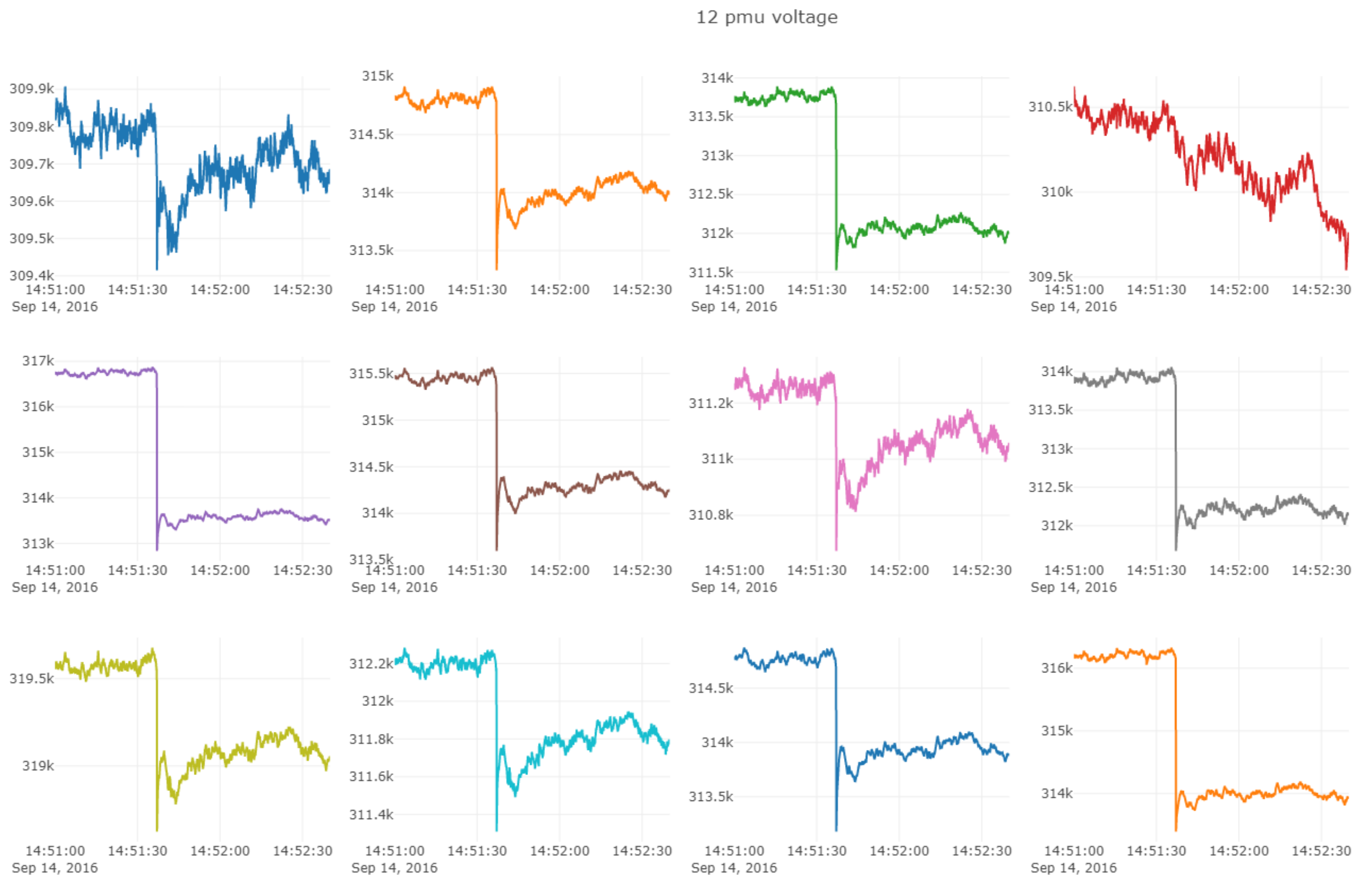
Examples of Detected events

- Voltage event

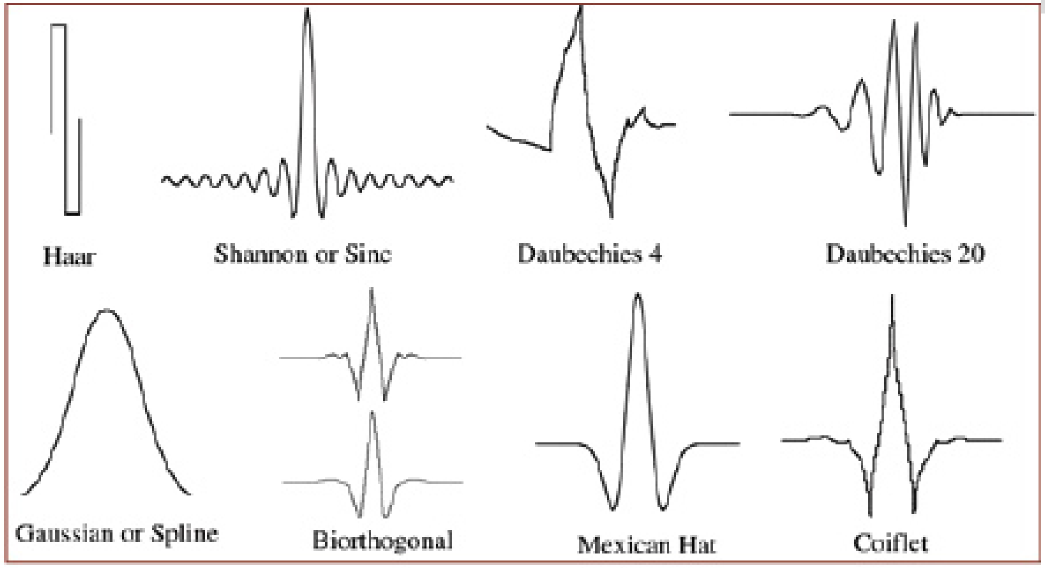


Examples of Detected events

- Voltage event



Wavelet analysis



"The wavelet transform is a tool that cuts up data, functions or operators into different frequency components, and then studies each component with a resolution matched to its scale"

---- Dr. Ingrid Daubechies, Lucent, Princeton U

- Wavelets transform:
  - Use small waves, so called wavelets, to provide localized time-frequency analysis.
  - Scaling (*stretching/compressing it; frequency band*) and shifting (*delaying/hastening its onset*) original waveform
    - Low scale→ compressed wavelets→ high frequency
    - High scale→ stretched wavelets→ low frequency
  - Assign a coefficient of similarity
- Benefit for the non-stationarity signals

Offline Anomaly Detection based on Wavelet Analysis

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*

PMU Signals → Wavelet-based multi-resolution analysis (MRA) → Wavelet coefficients extraction → Moving window based anomaly candidates → Anomaly scoring → Spatiotemporal correlation analysis → Final list of detected anomalies

PMU Signals

Voltage phasors

Frequency

ROCOF

Wavelet-based multi-resolution analysis (MRA)

Signals are transformed into time-frequency domain

MRA decomposes signals into approximation (A) and detail (D) components

Wavelet coefficients extraction

Extract the detail wavelet coefficients at scales matching the durations of events

Moving window based anomaly candidates

Identify anomaly candidates at multiresolution levels with moving-window-based anomaly detection

5-min bandwidth of moving window

Anomaly scoring

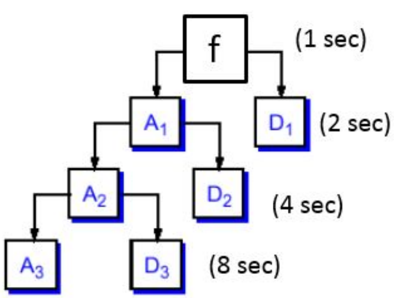
Anomalous score is set to be 1 when an anomaly is detected at each resolution level.

The anomaly score matrices are the summation of scores at multi-resolution levels across all PMUs

Spatiotemporal correlation analysis

False alarms or localized events corresponds to relatively weak spatiotemporal correlations

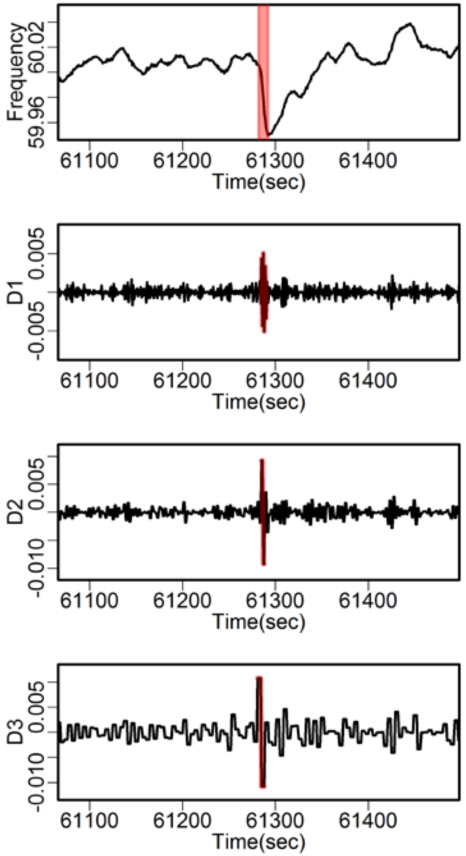
Final list of detected anomalies



MRA decomposition Tree

Anomaly Scoring and Verification

- The anomaly score matrices were calculated across 12 PMUs at multiresolution levels for each PMU attribute.
- Red line shows a historical recorded event at each multi-resolution level

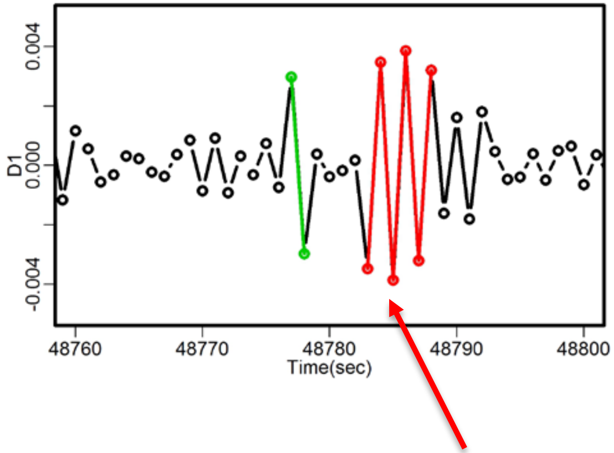


(a) Frequency signal

(b) MRA wavelet coefficient at D1;

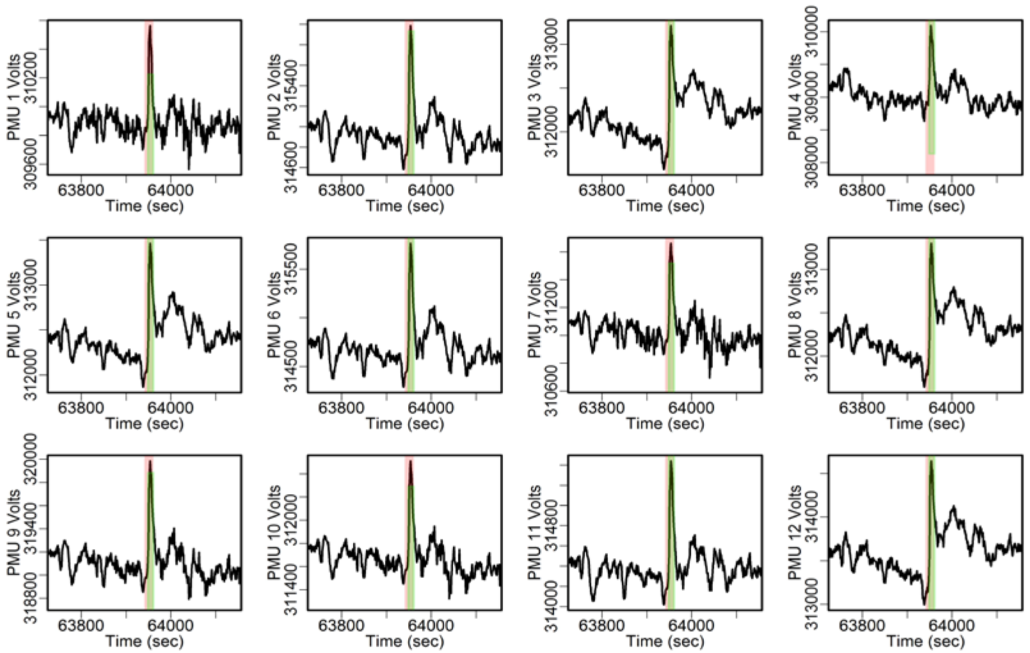
(c) MRA wavelet coefficient at D2;

(d) MRA wavelet coefficient at D3.



More than 3 sequential points exceeded the threshold and counted as an event. +1 added to the anomaly score matrices.

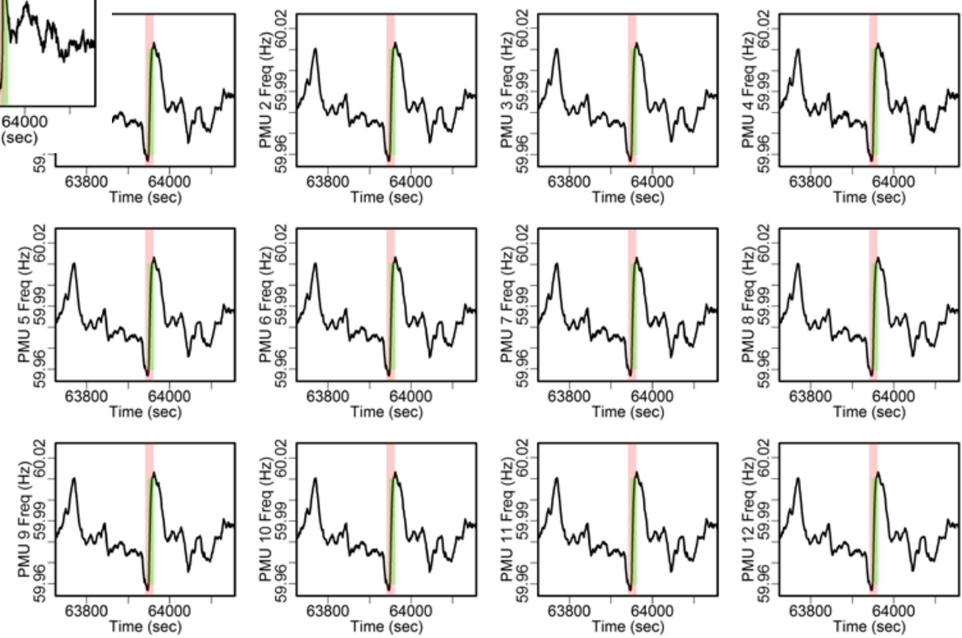
Examples of Detected Anomalies (1)



An example of detected PMU voltage anomaly where the PMUs have consistent behaviors and strong cross-correlations.

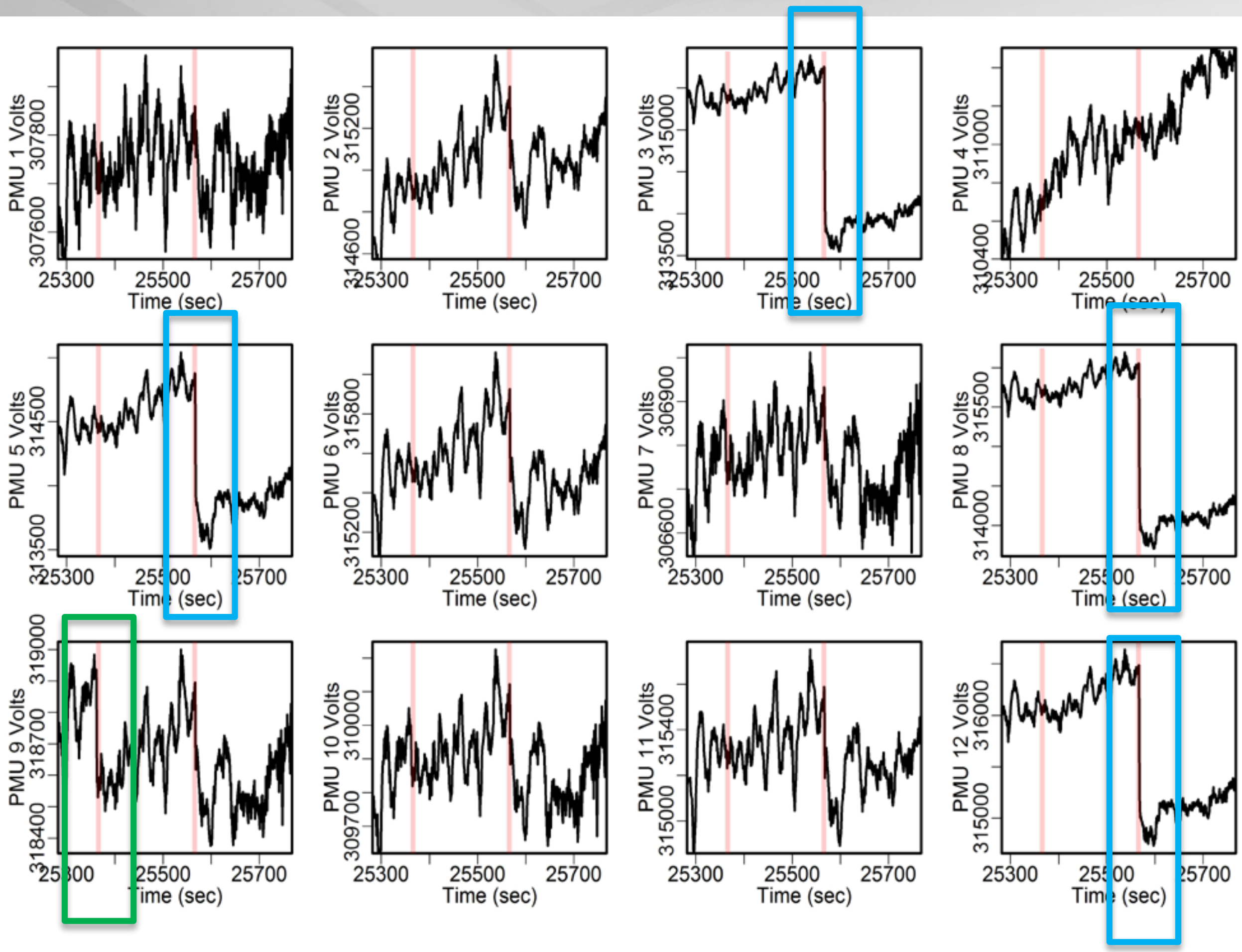
Red marks: detected events

Green marks: recorded historical events by NERC



An example of detected PMU frequency anomaly where the PMUs have consistent behaviors and strong cross-correlations.

Examples of Detected Anomalies (2)

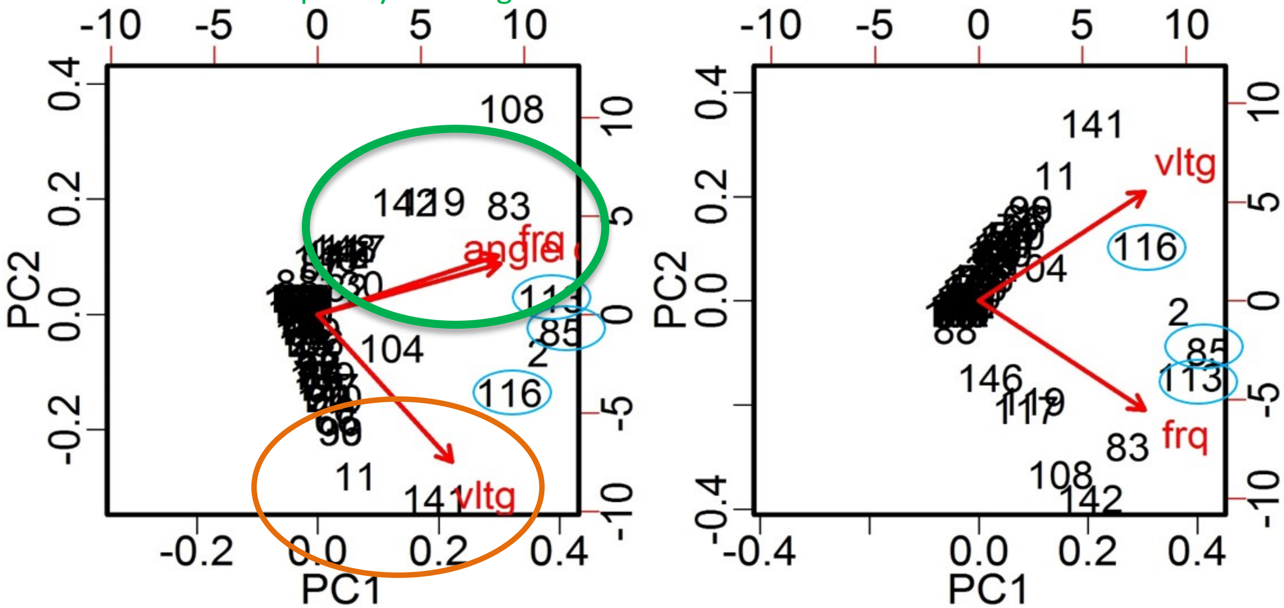


Example of voltage event detected at different local units. The detected events for each unit are marked in red.

- Local anomalies
- 7 out of 12 units did not evidence the same anomalies.
- The first event occurred at unit 9 only
- The second event happened at units 3, 5, 8 and 10, respectively.

Principal Component Analysis (PCA)

Strong anomalies in both frequency and angle variation



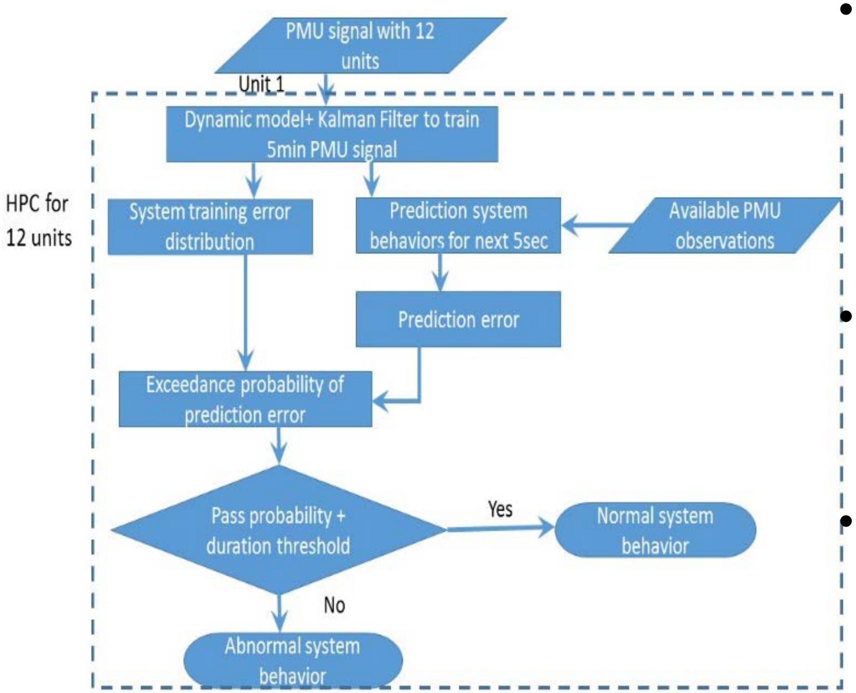
outstanding voltages anomalies

The left panel shows the first two principal components of three attributes (voltage, angle and frequency).

The right panel shows the PCA by removing the redundant angle variation. The voltage and frequency are nearly orthogonal factors

PCA Biplots of detected events using different PMU attributes. The historical recorded events are circled in blue.

Online Anomaly Detection Based on Dynamic Machine Learning



Flow chart of online detection framework for PMU measurements.

- The second order polynomial dynamic regression model is built sequentially for PMU of subsequent 5-minute time windows. Kalman filter is applied to compute filtered values of the state vectors, together with their covariance matrices.

- The training and prediction errors are obtained by model fitting and short-time prediction using available PMU observations.

- For the short-term predictions, we assume that the prediction errors and the training errors follow the same distributions. The cumulative probability distribution (CDF) of prediction errors is approximated to be normal and characterized by the mean and variance of the training errors.

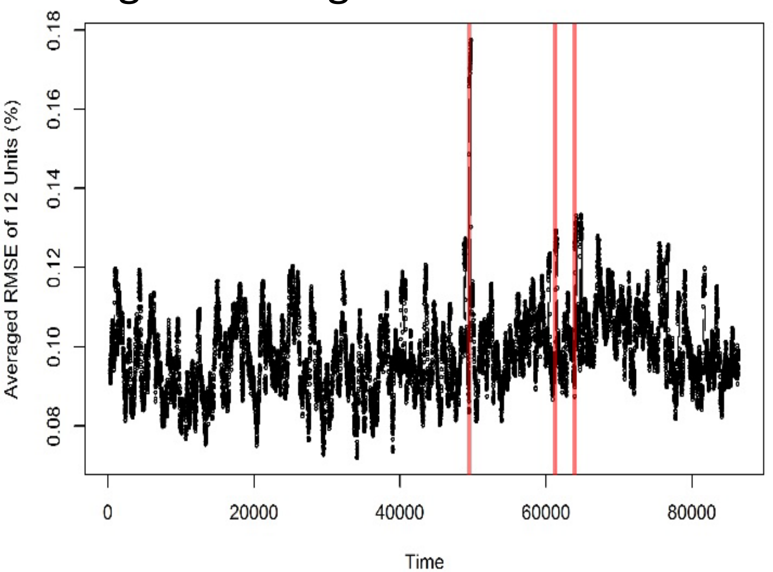
- A threshold of P_i can be used to screen the anomaly candidate points in the PMU data, based on whether its corresponding exceedance probability is greater than the threshold.

$$P_i(X \leq x) = \max(P_i(X \leq x), 1 - P_i(X \leq x))$$

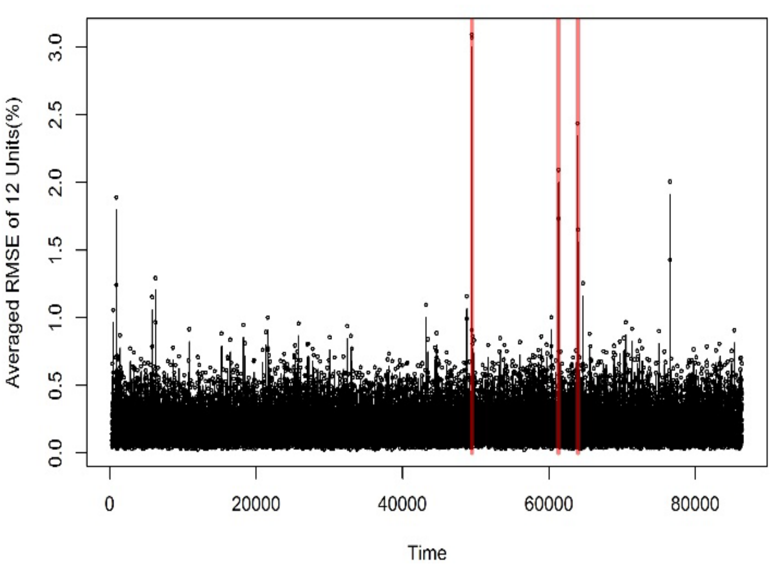
Dynamic Model Evaluation: Root Mean Square Error(RMSE)

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Averaged training RMSE across 12 Units



Averaged prediction RMSE across 12 Units



The red vertical lines show temporal locations of recorded events

Training RMSE:

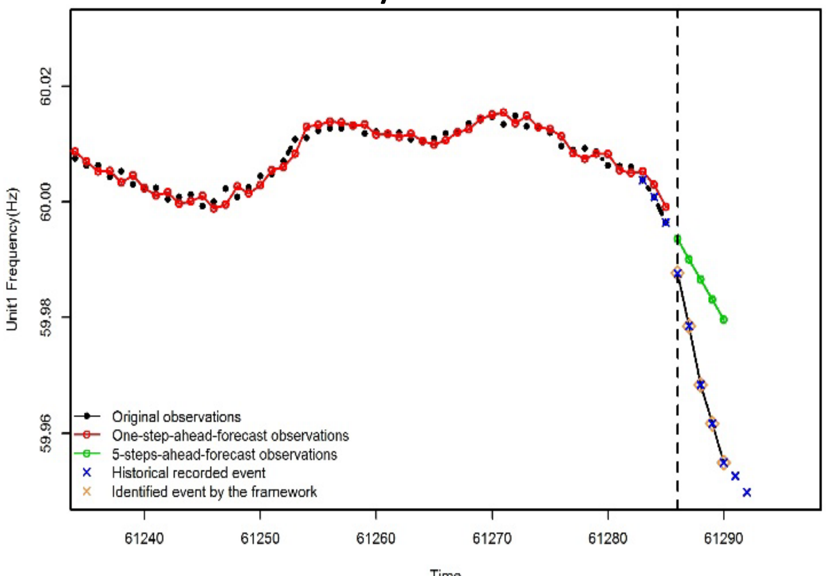
- RMSEs shows the satisfactory goodness of fit of the dynamic model.
- RMSEs are generally under 0.12% for the non-events time period.
- RMSEs increase slightly during the actual events occurred

Prediction RMSE:

- RMSEs shows the accurate predictions of next 5-sec
- RMSEs over 1.5% are highly likely to have some abnormal system behaviors
- RMSEs are relatively high (>2%) for the historical recorded events periods

Example of Event Detected and Detection Rate

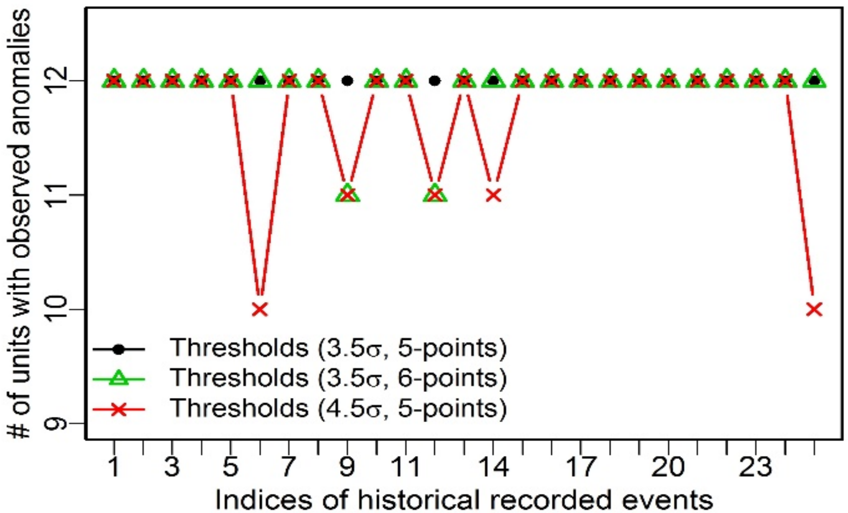
Historical recorded event and anomaly event detected by the framework



- For such an actual event, the deviations or relative errors increase with the time into the events

- The exceedance probability of the relative errors and the duration are compared to the thresholds to confirm anomalies.

The detection rates of historical recorded events



- 28-day PMU data with 25 historical recorded events are used to evaluate the framework
- Detection rates are calculated for different combinations of probability and duration threshold
- The optimal thresholds setting:
  - ✓ exceedance probability threshold is 3.5σ (i.e., the prediction error is beyond 3.5 times of the corresponding standard deviation σ).
  - ✓ duration threshold is 5-points (i.e., seconds), which means at least 5 sequential points need to pass the screening in order to confirm an event

Preliminary Results

- Spark cluster for ML and PMU (big data) analysis was deployed. It is based on the PNNL institution cloud system
- PMU data has been collecting in PDAT format (PMU data stream from PBA to PNNL EIOC)
- Methodologies for both online and offline anomaly detection have been developed
  - Enhanced robustness to bad data
- Python (PySpark) modules are under development
  - PDAT data extraction
  - Event detection (based on thresholds)
  - Wavelet anomaly detection
  - Dynamic nonlinear model and Kalman filter based online detection framework