Parametric Data Analysis - Pre-empting Link Failures

Granthana Rangaswamy
Arushi Sharma
Anju John
Matt Bergeron
Venkat Ramesh

Meta

Introduction - Why Parametric Analysis is needed?

Background

- Next generation AI systems will see an increased impact from SerDes related issues due to shrinking margins, increasing speeds and increasing complexity.

- With higher speeds, more complex modulation schemes, reduced signal-to-noise ratios (SNRs), and longer job runs for AI clusters, high-speed designs are even more critical and must be designed with utmost care.

- Meta's 24K AI clusters will have thousands of SerDes running at 112Gbps and beyond and cannot afford to have unplanned resource unavailability and job restarts.

Meta

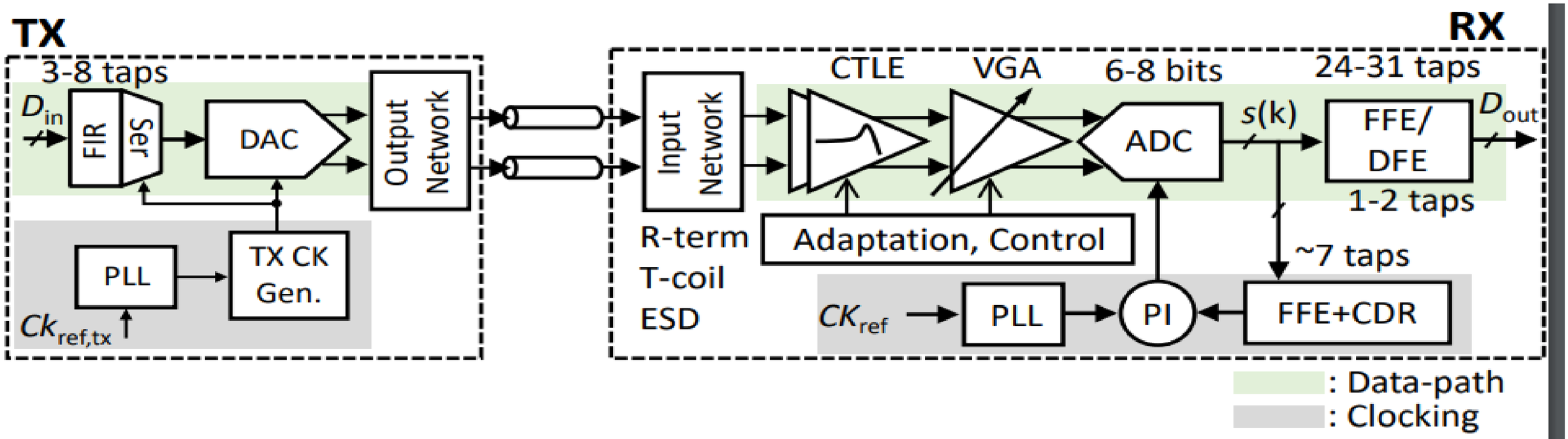
Introduction - Why Parametric Analysis is needed?

Goal

- Parametric data analysis is a framework in which we collect, parse and analyze SerDes data on new high speed systems.
- This automated collection and analysis of Ethernet SerDes parameters can be used to build anomaly detectors to predict link-failures based on large-scale datasets.
- These anomaly detectors, enhanced with machine learning algorithms and refined pass/fail criteria, will enable preemptive detection of link issues and shifts in margin distributions.
- This capability will accelerate the deployment and effective management of next-generation systems.

Meta

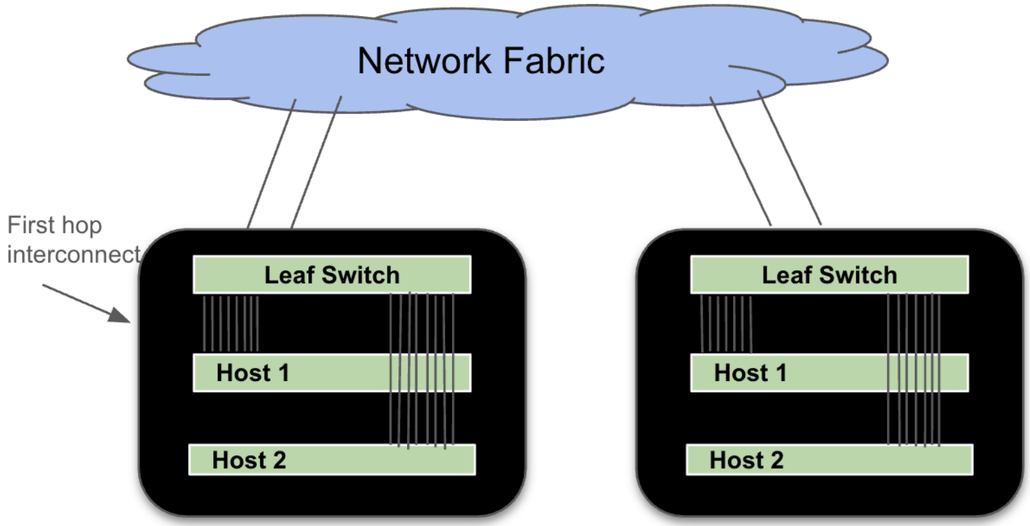
SerDes and its dependency on AI clusters

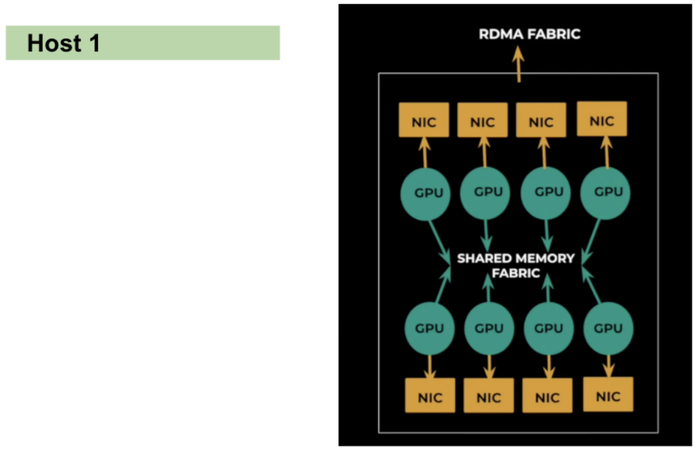


- High speed fabric channels are typically point to point.
- The channel behaves like a low pass filter attenuating the high frequency components.
- Compensating for inter symbol interference (ISI) is critical and is mainly done in digital domain.
- Above picture is a sample of the DAC/ADC-based architecture and numerous methods that can be used to equalize a lossy channel.

Pic: Circuits and Systems for Signal Processing (CASSP) Lab.

AI BackEnd Network





END-POINTS

- 8x GPU and 8 NICs per server connected by PCIe
- Within server: Shared memory fabric
- Across servers: RDMA fabric

Goal

- Quantify the impact of network interface flaps/related failures to system reliability in the downlinks
- Interface flap is an event where the link transitions from Up to Down to Up
- The flap interval can last several ms to seconds
- Packet drops occur during link flaps, result in application being timed out
- Factors for link flap
  - Bad signal integrity
  - Bad cables, NIC
  - Failure in the first hop upstream device
- Repair Actions
  - Reseat, Swap endpoints

Parametric analysis implementation

- APIs to invoke the collection of data.
- Determine the parameters, and time intervals to collect the data.
- Integrate with vendor ASIC SDK.
- Collect metrics periodically.

- Parsing the log files
- Normalizing the schema
- Build json file

- Understand Input to Output correlation
- Predetermined masks for serdes parameters
- Serdes stats acceptable range
- Deploy anomaly detectors

- Fleet analysis and utilize the data to determine next steps.
- On the fly system evaluation

Types of Parametric Data that can be collected

Per-SerDes Data Collection – some examples

- BER with PRBS (disruptive)
- DFE Settings
- DAC Control Values
- VGA Settings
- Rx DFE Taps
- FEC Histogram

System Level Data Collection

- Dynamic information such as temperature, voltage: things which can directly affect SerDes performance captured at periodic intervals
- Static System information such as link speed, PAM4 vs. NRZ setting, reach-mode, FW, auto link training should be captured during manufacturing, at first boot in DC and after system maintenance.

Methodology for Parametric Data Analysis

- Due to adaptation variation, connector or manufacturing can clutter the data.
- Rx tuned parameters (CTLE, DFE taps) is used as criteria for data cleaning.

Data Cleaning

- Training parameters (theta, cost gradient) can be computed based on the existing measurement data
- Any new data can utilize these training parameters

Is new data clean?

No

Collect data again

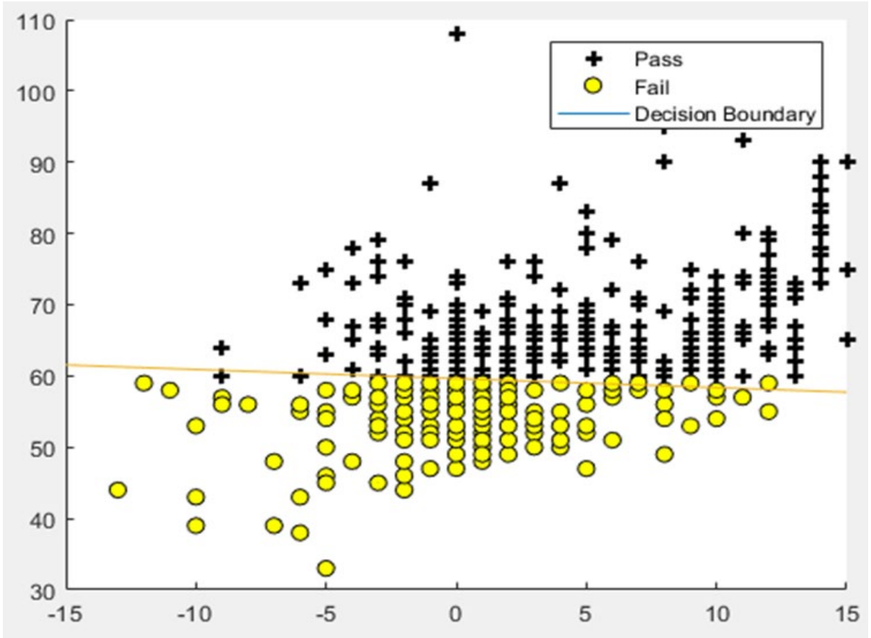
Yes

Apply Logistic / Random Forest Algorithm on new Data

Pass or Fail?

Passed with X% accuracy

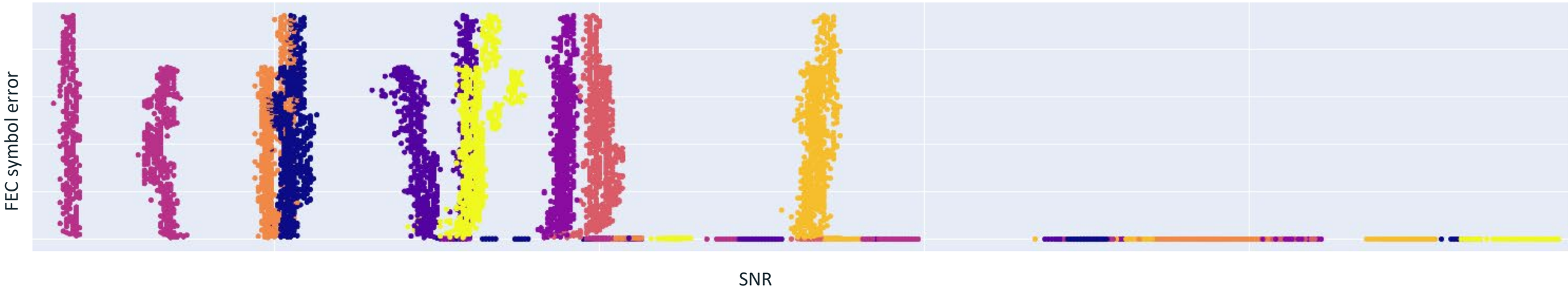
Failed with 1-X% accuracy



At Scale Data Analytics

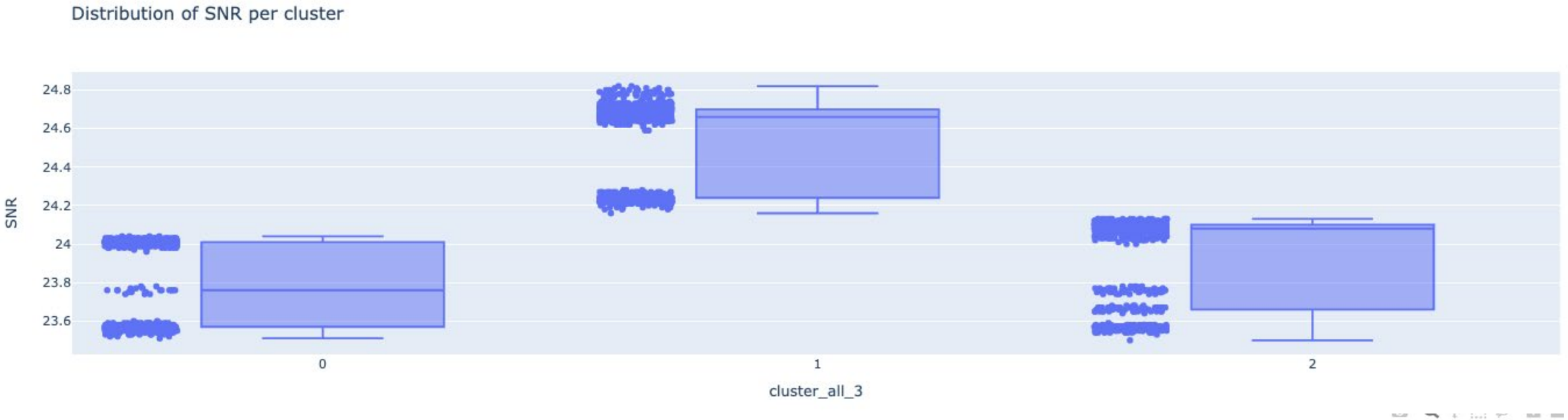
Each channel/lane has a fundamental “good enough” parameters error cliff

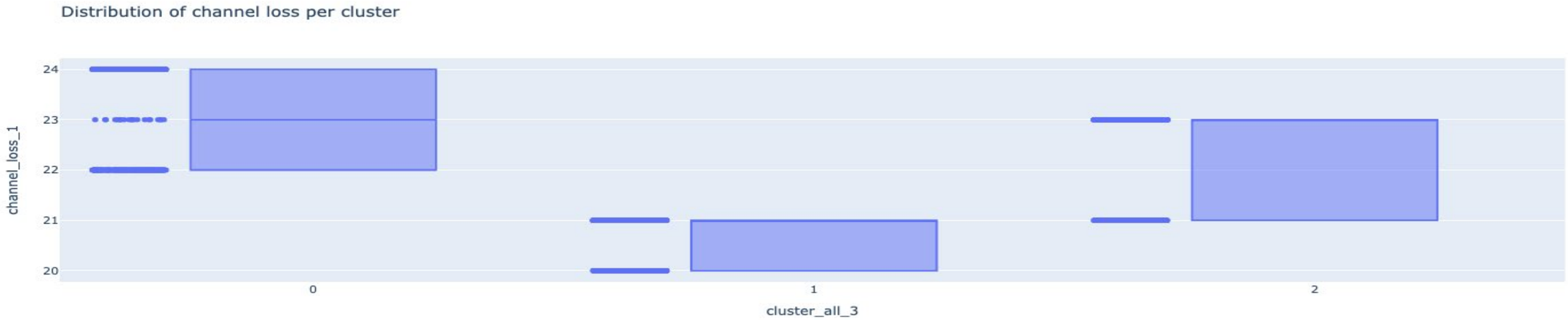
- Inform threshold detection
- Predict link-flaps



Requires FEC histograms and other parametrics
Plot above is measured on Meta built system

At Scale Data Analytics





Call to Action

- *Engage with OCP Test & Validation to align on a common set of requirements across all high speed interfaces*
  - *Identify a common list of parameters which will be abstracted*
  - *Identify a common API/method to fetch this information*

- *Ensure that your product roadmap complies with the requirements*

- *Work closely with Meta on developing validation and testing infrastructure that reduces the friction during AI infrastructure deployment and maintenance*


Meta