ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-MOPAB288

REAL-TIME EDGE AI FOR DISTRIBUTED SYSTEMS (READS): PROGRESS ON BEAM LOSS DE-BLENDING FOR THE FERMILAB MAIN INJECTOR AND RECYCLER

K. J. Hazelwood*, M. R. Austin, M. A. Ibrahim, V. P. Nagaslaev, D. J. Nicklaus, A. L. Saewert, B. A. Schupbach*, K. Seiya, R. M. Thurman-Keup, N. V. Tran, A. Narayanan[1]
Fermi National Accelerator Laboratory†, Batavia, IL, USA
H. Liu, S. Memik, R. Shi*, M. Thieme*
Northwestern University‡, Evanston, IL, USA
[1]also at Northern Illinois University, DeKalb, IL, USA

Abstract

The Fermilab Main Injector enclosure houses two accelerators, the Main Injector and Recycler. During normal operation, high intensity proton beams exist simultaneously in both. The two accelerators share the same beam loss monitors (BLM) and monitoring system. Beam losses in the Main Injector enclosure are monitored for tuning the accelerators and machine protection. Losses are currently attributed to a specific machine based on timing. However, this method alone is insufficient and often inaccurate, resulting in more difficult machine tuning and unnecessary machine downtime. Machine experts can often distinguish the correct source of beam loss. This suggests a machine learning (ML) model may be producible to help de-blend losses between machines. Work is underway as part of the Fermilab Real-time Edge AI for Distributed Systems Project (READS) to develop a ML empowered system that collects streamed BLM data and additional machine readings to infer in real-time, which machine generated beam loss.

READS OVERVIEW

The Real-time Edge AI for Distributed Systems (READS) project is a collaboration between the Fermilab Accelerator Division and Northwestern University. The project has two objectives; first to implement Machine Learning (ML) into the future Delivery Ring slow spill regulation system [2] for the Mu2e experiment [3, 4], and second to create a real-time beam loss de-blending system for the Main Injector (MI) accelerator enclosure also utilizing ML [5].

Beam Loss De-blending

The Main Injector and Recycler Ring (RR) accelerators share a tunnel and one beam loss monitor (BLM) system. The 8 GeV permanent magnet Recycler was originally built as an anti-proton storage ring for the Tevatron collider [6]. Anti-proton losses in Recycler were insignificant compared to the 8 GeV to 120 GeV proton losses from Main Injector; there was less need to monitor ionization beam losses from Recycler. When the Tevatron was decommissioned, Recycler was re-purposed as a proton stacker for Main Injector 120 GeV NuMI beam operation [7] as well as for 8 GeV Muon g-2 experiment beam delivery [8]. Currently, normal operation of the accelerator complex has high intensity beams in both Main Injector and Recycler simultaneously. Beam losses from both machines are now a large concern. The origin of loss on any of the 259 operational BLMs can be difficult to attribute to any one machine. However, experts can often attribute losses to either Main Injector or Recycler based on time (Fig. 1), machine state, and location (Fig. 2).

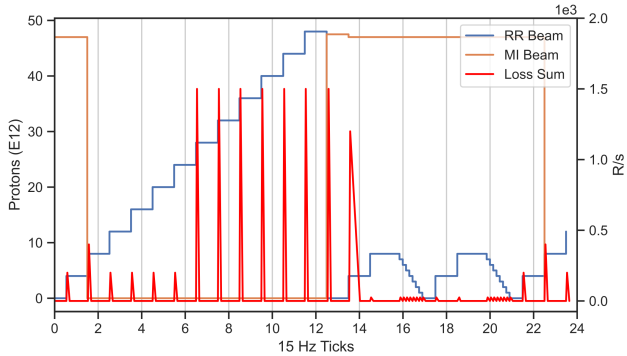

Figure 1: Example illustration of overlapping beam events and losses in the MI and RR accelerators.

Using streamed distributed readings and real-time ML inference hardware, this project aims to replicate and improve upon the machine expert's ability to de-blend losses between machines.

PIRATE CARD DEVELOPMENT

In order to satisfy the data requirements for this project, a parasitic VME bus reader card, commonly referred to as a Pirate Card, is being developed and integrated into the existing MI BLM system. Each of the 7 BLM nodes distributed around the 2.2 mile Main Injector complex consists of a VME Crate Processor, Control Card, Timing Card, and an array of digitizers [9]. The sole responsibility of the Pirate Cards is to intercept the BLM values of each digitizer throughout the beam cycle without disturbing normal operations of the system. The digitized BLM values will be

* Equal contribution
† Operated by Fermi Research Alliance, LLC under Contract No.De-AC02-07CH11359 with the United States Department of Energy. Additional funding provided by Grant Award No. LAB 20-2261 [1].
‡ Performed at Northwestern with support from the Departments of Computer Science and Electrical and Computer Engineering.

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-MOPAB288

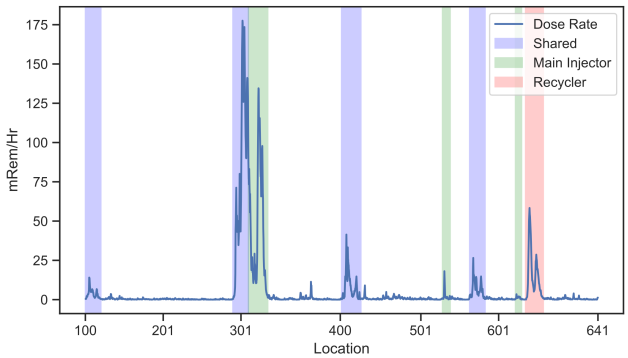

Figure 2: Location dependency of MI and RR beam loss as seen from tunnel residual dose rates.

packaged up with other relevant machine data like beam intensities, cycle events, and MI momentum. This datagram will be sent over the network for ML model training data as well as for the eventual FPGA ML model implementation. The Pirate Cards have been designed and are currently being manufactured. Delivery of the cards is expected by late spring 2021; in time to collect higher frequency training data to be used in training a final ML model.

DATASETS

A **Sample Dataset** is being continuously generated using actual accelerator operations data. The data in the Sample Dataset is the same in structure and shape as the data that is expected to be used for training the final ML models, albeit at a much reduced frequency (15 Hz) compared to data rates expected from the Pirate Cards. Data rates for the Sample Dataset are constrained by limitations of the existing BLM system and ACNET controls network. An algorithm has been created to label the BLM losses by machine when possible. This data will help inform the extent and structure of data coming from the Pirate Cards. Initial ML models are being designed and trained from this dataset.

The **Training Dataset** will be created using data streamed from the aforementioned Pirate Cards. The expected Training Dataset rate is 333 Hz which corresponds to the current rate at which the BLM system nodes poll their digitizers for new BLM sums. Normal operation of the accelerators does not allow for much opportunity to record BLM readings when beam exist in only one machine. For this reason, fairly involved studies have been requested in late spring 2021, just before the Fermilab accelerator complex has it's annual maintenance shutdown. The study will involve manipulating the beam event time line to purposefully keep events for Recycler and Main Injector from overlapping, thus only having beam in one machine at a time. This will allow for the proper attribution and labeling of beam losses to one machine using the same algorithm used for the Sample Dataset. To ensure that a broad range of loss conditions are captured for the Training Dataset, moderate beam losses will be generated at all locations in both machines using various miss-configurations of the machines.

MODEL ARCHITECTURE

The De-Blending Network (DBLN) is comprised of three parts: a BLM network, a State Network, and an Aggregation Network (Fig. 3). At each point in time, the DBLN maps observations of the last n BLM loss signatures $l_n \in \mathbb{R}^{n\times 259}$ and machine data $e_n \in \mathbb{R}^{n\times 9}$ to class-conditional probabilities over individual accelerators *per BLM*: $p(a_i|l_n, e_n) \in \mathbb{R}^{259\times 2}$. Note here the overloading of the term "loss": when referring to the BLM losses we use l, and when referring to the mathematical quantity related to the ML model performance we use ℓ.

The **BLM Network** is a convolutional neural network (CNN) with two max-pooled convolutional layers followed by two linear layers. The BLM network learns a mapping between the raw BLM loss data $l_n \in \mathbb{R}^{n\times 259} \rightarrow B_n \in \mathbb{R}^k$. This vector B_n is then ingested by the Aggregator where it is conditioned on a representation of the machine state generated by the State Network.

The **State Network** is a two layer fully-connected multilayer perceptron (MLP) that learns a mapping between the last n state observations $e_n \in \mathbb{R}^{n\times 9} \rightarrow S_n \in \mathbb{R}^k$. The output S_n serves as a conditioning mechanism for the representation of the BLM loss signature B_n.

The **Aggregator Network** is a three layer fully-connected MLP that learns to map $B_n \oplus S_n \in \mathbb{R}^k$, where $\oplus$ is the elementwise sum, to class-conditional probabilities over $p(a_i|l_n, s_n) \in \mathbb{R}^{259\times 2}$. We choose the elementwise sum instead of concatenation to make the model more compact.

To train the model, we use Binary Cross-Entropy Loss and the Adam Optimizer with learning rate = 0.001. The final model has 1.3M trainable parameters.

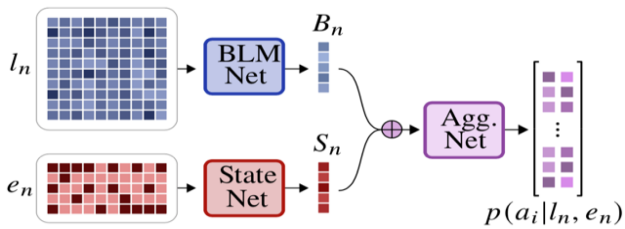

Figure 3: DBLN model architecture.

PRELIMINARY RESULTS

Initial results using the Sample Dataset show promising performance. Figure 4 details training accuracy and loss over the first 1000 batches using the past $n = 2$ observations and batch size = 32.

Figure 5 (A) shows the measured beam intensities in MI and Recycler over 24, 15 Hz ticks. Section (B) shows the normalized BLM measurements at each tick. Sections (C) and (D) show the output of our model scaled by the BLM loss intensity (significance). From these plots, we can see that our model is appropriately classifying the losses in each region. As beam is extracted from Recycler to MI; our model recognizes the change in the loss signature and switches the inferred label from RR to MI in turn.

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-MOPAB288

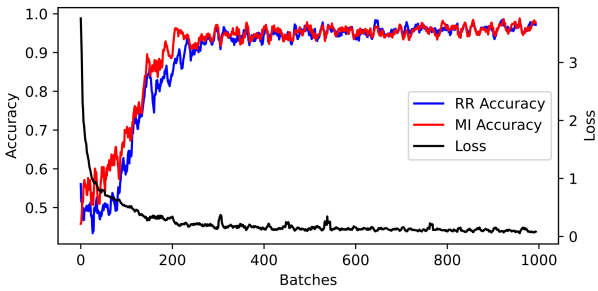

Figure 4: RR and MI training accuracy over 1000 batches.

Overall, high validation accuracies (95% and 96% for MI and RR respectively) is evidence that our model is learning meaningful mappings between BLM loss profiles and their machine of origin.

Model Confidence

Of particular interest is the model's behavior on the BLM losses which cannot be attributed to a single machine, i.e. loss profiles acquired when MI and RR operate simultaneously. Each row in the output $p(a_i|l_n, s_n) \in \mathbb{R}^{259\times2}$, corresponds to $[p(\text{MI}), p(\text{RR})]$. Probabilities ≥ 0.5 are treated as positive identifications and < 0.5 as negative identifications. Performing inference on these data with unknown labels yields the confusion matrix in Table 1, where MI/RR (+) and MI/RR (-) represent positive and negative identifications, respectively.

Table 1: Model Confusion Matrix

	MI (+)	**MI (-)**
RR (+)	1%	2%
RR (-)	77%	19%

Presently, the model is disproportionately recognizing these unknown losses as originating from MI and *not* RR. Experimentation is underway to better understand the model's behavior on these data with unknown labels.

MODEL IMPLEMENTATION

The ML model will be implemented as an IP core on the Intel Arria-10 SOC. The board contains a FPGA side and a hard processor subsystem (HPS) side which have fast communication through HPS-FPGA bridges between them. The FPGA side can be used to implement the DBLN network for processing the data. The HPS side has Ethernet and can do complicated calculations. The HPS will also be useful for updating the implemented DBLN network by partial reconfiguration.

The hls4ml+Quartus tool flow will be used to generate an initial design for the ML IP. Based on the model development described in previous sections, the saved model will act as an input to hls4ml and will generate an implementation in C++ which uses HLS Compiler for hardware design. The

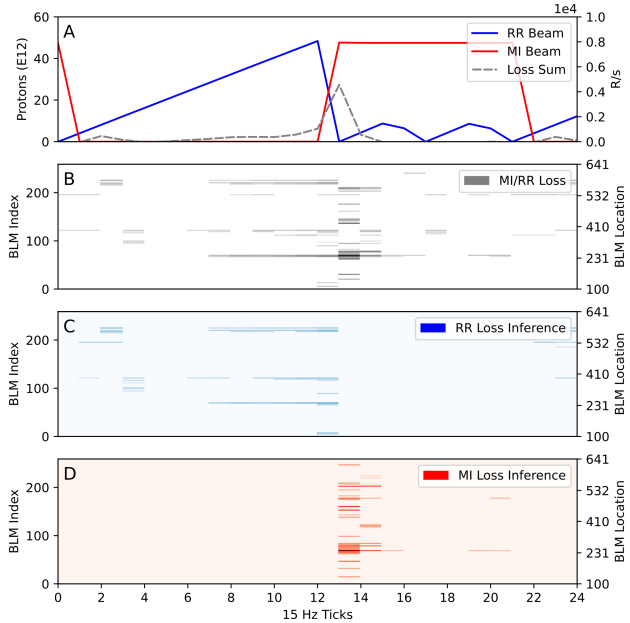

Figure 5: Model inference on a single beam extraction from RR to MI using Sample Dataset.

Quartus backend of hls4ml will be used as a starting point with additional customization done later. Finally, The DBLN network will be implemented as an IP core and connected to HPS using the Platform designer.

During the design phase of the IP, there must be trade-offs between the expected latency and the limited resources. Various methods can be adopted to optimize the implementation by exploiting pipeline and parallelism. For example, unrolling the potential loops, modifying the initial interval and adding registers between each layer can all be used to help achieve the desired data processing pipeline. Multiple parallel kernels will be used, this will likely translate into parallel data paths. To fully utilize the limited resources on the FPGA, with respect to constraints imposed by other functionality that is co-located on it, we need to carefully select a proper precision of the data representation and consider reuse of the implemented kernel and BRAM buffers.

SUMMARY

The READS Main Injector accelerator enclosure beam loss de-blending project is progressing well. A Sample Dataset has been generated using MI/RR readings and a very promising preliminary ML model has been created from the data. To meet the project's data needs, a custom BLM node VME bus reader card, commonly referred to as a Pirate Card, has been designed and is being manufactured. Delivery of the Pirate Cards is expected late spring 2021, just in time to collect high fidelity training data before the planned Fermilab accelerator complex summer maintenance shutdown. Hardware for the final FPGA ML model implementation has been acquired and is being developed on.

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-MOPAB288

REFERENCES

[1] "Data, Artificial Intelligence, and Machine Learning at DOE Scientific User Facilities", Department of Energy National Laboratories, USA, Rep. LAB 20-2261, 2020.

[2] A. Narayanan *et al.*, "Optimizing Mu2e Spill Regulation System Algorithms", presented at the 12th Int. Particle Accelerator Conf. (IPAC'21), Campinas, Brazil, May 2021, paper THPAB243, this conference.

[3] L. Bartoszek *et al.*, "Mu2e Technical Design Report", FERMILAB, Batavia, IL, USA, Rep. FERMILAB-TM-2594, 2014.

[4] V. Nagaslaev *et al.*, "Third Interger Resonance Slow Extraction Using RFKO at High Space Charge", FERMILAB, Batavia, IL, USA, Rep. FERMILAB-CONF-11-475-AD, 2011.

[5] K. Seiya *et al.*, "Accelerator Real-time Edge AI for Distributed Systems (READS) Proposal", 2021. arXiv:2103.03928

[6] G. Jackson, "The Fermilab Recycler Ring Technical Design Report. Revision 1.2", FERMILAB, Batavia, IL, USA, Rep. FNAL-TM-1991, 1996.

[7] R. Ainsworth *et al.*, "High Intensity Proton Stacking at Fermilab: 700 kW Running", in *Proc. 61st ICFA Advanced Beam Dynamics Workshop on High-Intensity and High-Brightness Hadron Beams (HB'18)*, Daejeon, Korea, Jun. 2018, pp. 136–140. doi:10.18429/JACoW-HB2018-TUA1WD04

[8] J. Grange *et al.*, "Muon (g-2) Technical Design Report", 2018. arXiv:1501.06858

[9] A. Baumbaugh *et al.*, "Beam Loss Monitor Upgrade User's Guide", Jul. 2010. https://beamdocs.fnal.gov/cgi-bin/sso/ShowDocument?docid=1410