Exploring FPGA in-storage computing for Supernova Burst detection in LArTPCs

Jovan Mitrevski, Benjamin Hawks, Tejin Cai, Pengfei Ding, Tom Junk, Kate Scholberg, Jieran Shen, Nhan Tran, Michael Wang, Tingjun Yang (Fermilab)

Motivation

Underground neutrino detectors can be used as supernova triggers for multi-messenger astronomy by providing pointing information to other observers [1], but to be effective, this data needs to be sent quickly, as can be demonstrated by Fig. 1, showing the time evolution of a supernova burst. Transferring all the data to the surface would take hours, while in-cavern power budgets are very limited: we explore using “in-storage computation” with FPGAs to provide pointing information more rapidly.

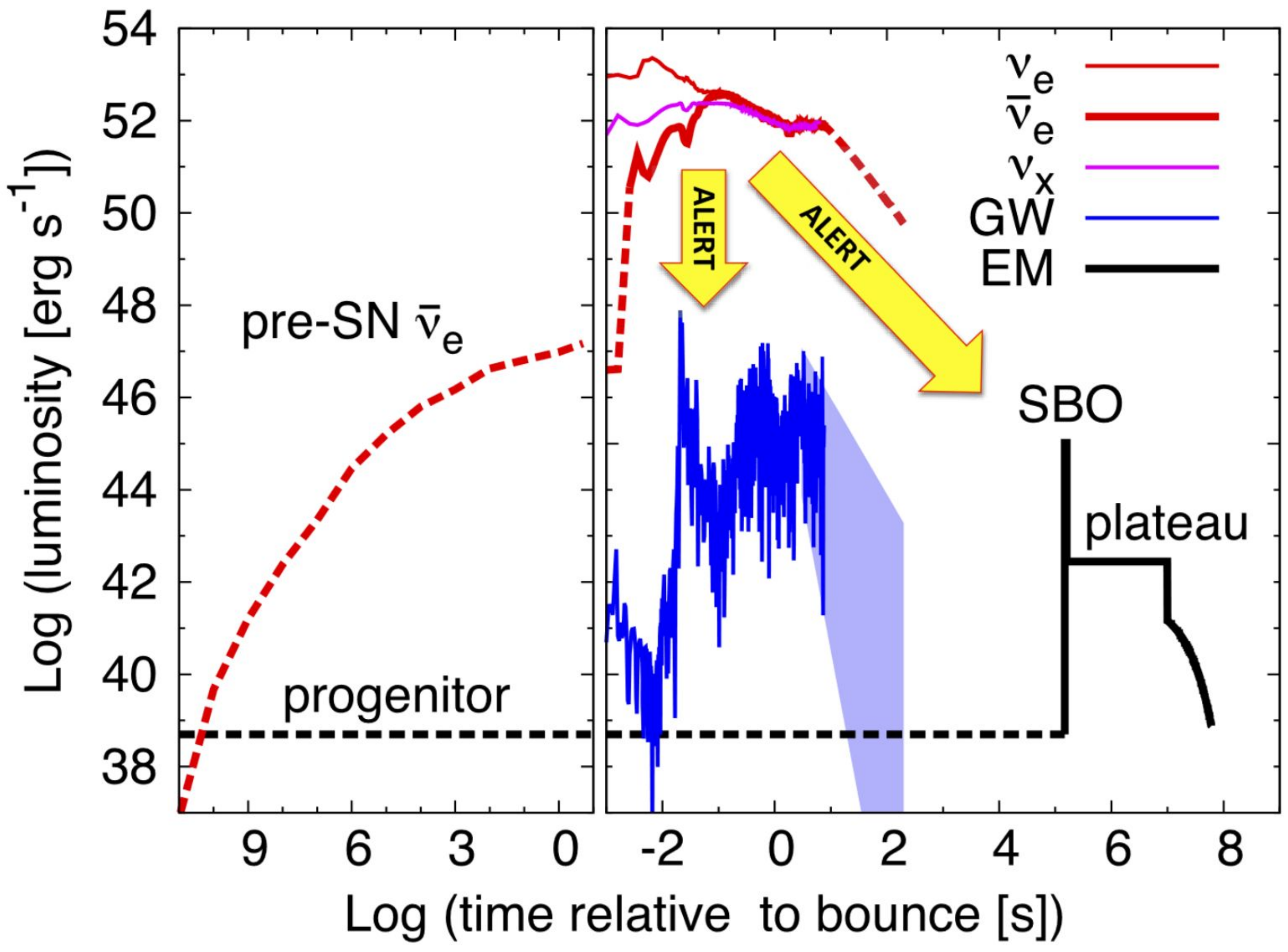

Figure 1: Time sequence for multi-messenger signals pre- (left panel) and post- (right panel) core collapse of a non-rotating 17M progenitor star. From: S Al Kharusi et al 2021 New J. Phys. 23 031201

Implementation & Algorithm

As a first proof of concept, we explore a simple task on the peer-to-peer accelerator scheme using an Alveo U55C FPGA with consumer NVMe SSDs. The POC algorithm/task being accelerated is a 1D CNN [2] to find regions of interest, as shown in Fig 2. To implement the task on the accelerator, the CNN is first quantized using QKeras [3] to use 4-bit weights and 5-bit activations, converted to HLS using hls4ml [4], then using Xilinx Vitis, the CNN is run as a kernel on the FPGA.

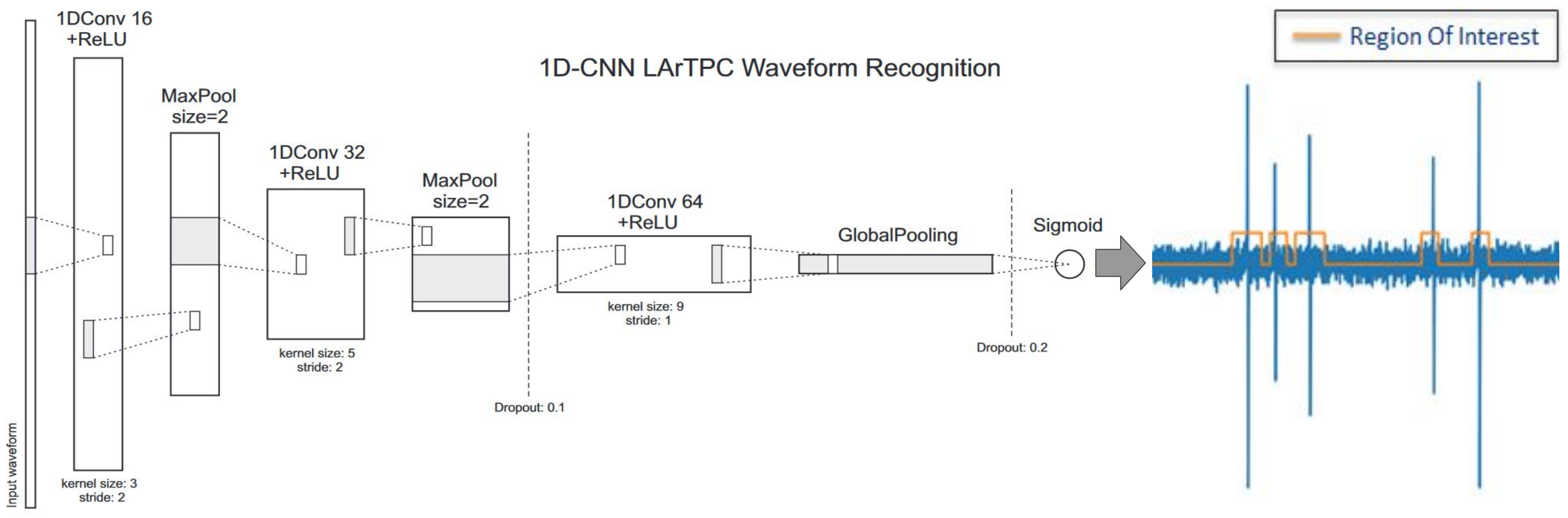

Figure 2: An example of a 1D CNN that can be used to find regions of interests for subsequent processing [2], either locally or on the surface.

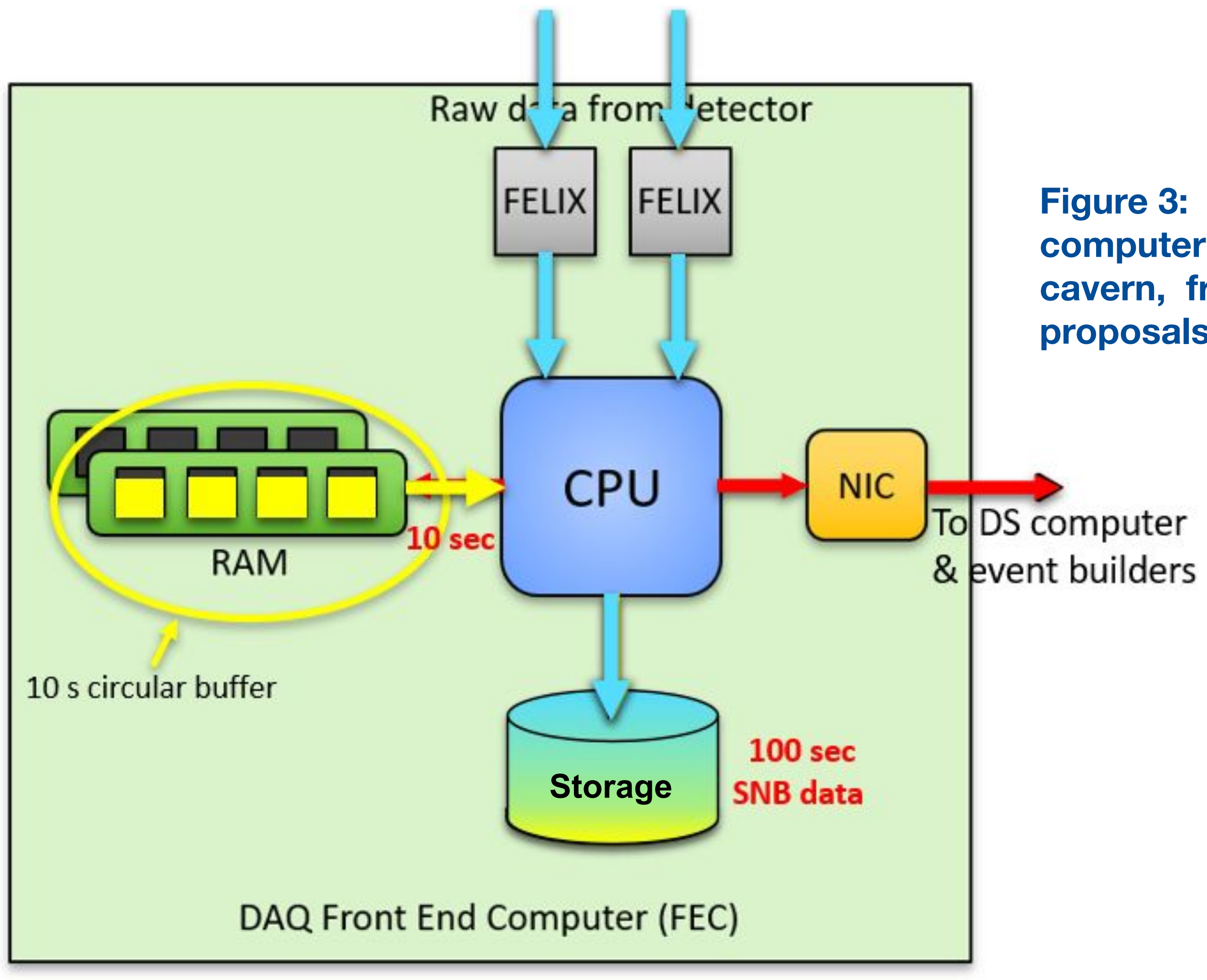

Figure 3: Example of a DAQ computer located in the cavern, from DUNE design proposals

Hardware

In an example of a DAQ computer located in a cavern, shown in Fig. 3, Our goal is to add processing to the storage in one of two possible schemes as shown in Fig. 4:

- Discrete accelerators (FPGA, GPU, etc.) that access the storage directly via peer-to-peer connections
- Unified storage and accelerator (FPGA) such as a “smart SSD”

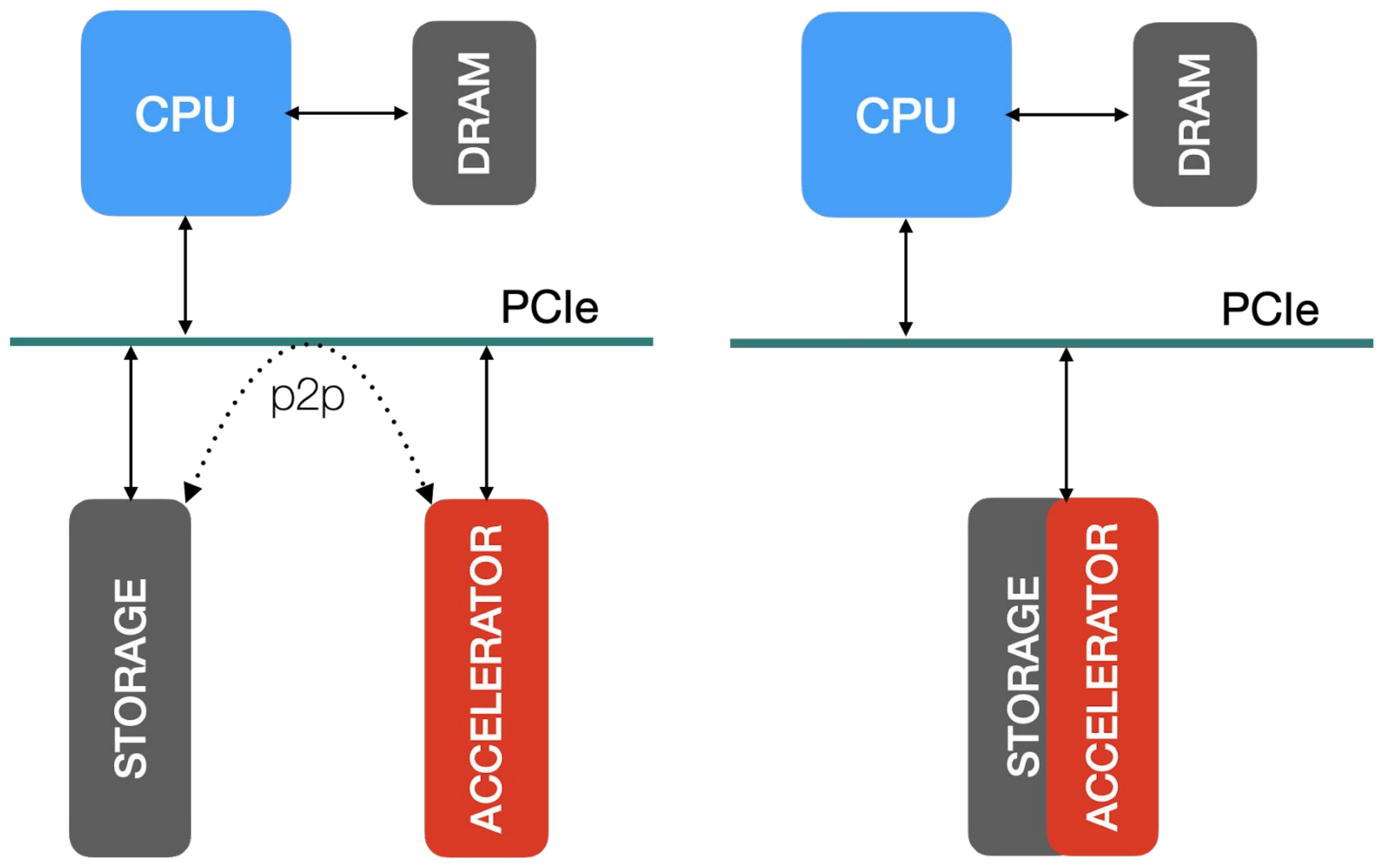

Figure 4: Examples of in-storage computing. The processing can be with a stand-alone accelerator accessing storage with peer-to-peer transactions, or it can be with an accelerator placed directly in the storage, i.e., a smartSSD.

Results & Conclusion

Preliminary results show significant speedup with low utilization on the Alveo U55C, indicating that using in-storage compute for more complex tasks may be useful.

- Applying the 1D CNN on a 1.7GB file located on an NVMe SSD has a 20× speedup on the Alveo U55C compared to serial code on an AMD EPYC 7313.
- Given the narrow bit widths used in the CNN, DSPs are not used, and only about 10% of the LUTs are used. As such, there is room for further processing on the Alveo.

References

[1] Abi et al 2020 JINST 15 T08008
[2] Uboldi et al., Nucl.Instrum.Meth. A1028:166371,2022
[3] Coelho Jr et al., Nat Mach Intell **3**, 675–686 (2021).
[4] Duarte et al, JINST 13 P07027 (2018).