FERMILAB-SLIDES-20-121-E

Fermilab

Design of a reconfigurable autoencoder algorithm for detector front-end ASICs

Christian Herwig

November 6, 2020

IEEE Nuclear Science Symposium

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

Motivation

- Detectors are becoming increasingly segmented
  - Response to more complex events (e.g. 'pileup' @LHC)
- Trigger challenge: how to manage trigger readout rates, while also benefitting from fine segmentation?
  - The key bottleneck is on-detector data reduction
- Traditionally, detector-specific ASICs simply "sum or sort", leaving the intensive processing to off-detector electronics
  - Meanwhile off-detector logic has become increasingly complex (tracking, clustering, event reco. on FPGAs)
- More computationally intensive on-detector processing may open avenues for enhanced trigger performance

Outline

- This talk presents a Neural Network (NN) autoencoder for front-end data compression on an ASIC, based on the CMS High-Granularity Endcap Calorimeter (HGCal).
- Our design seeks to:
  - ***enable*** more complex compression algorithms, with the potential to improve physics performance
  - ***customize*** the compression algorithm for individual sensors based on their location within the detector
  - ***adapt*** the compression algorithm for changing detector conditions (e.g. radiation damage, new beam configs)

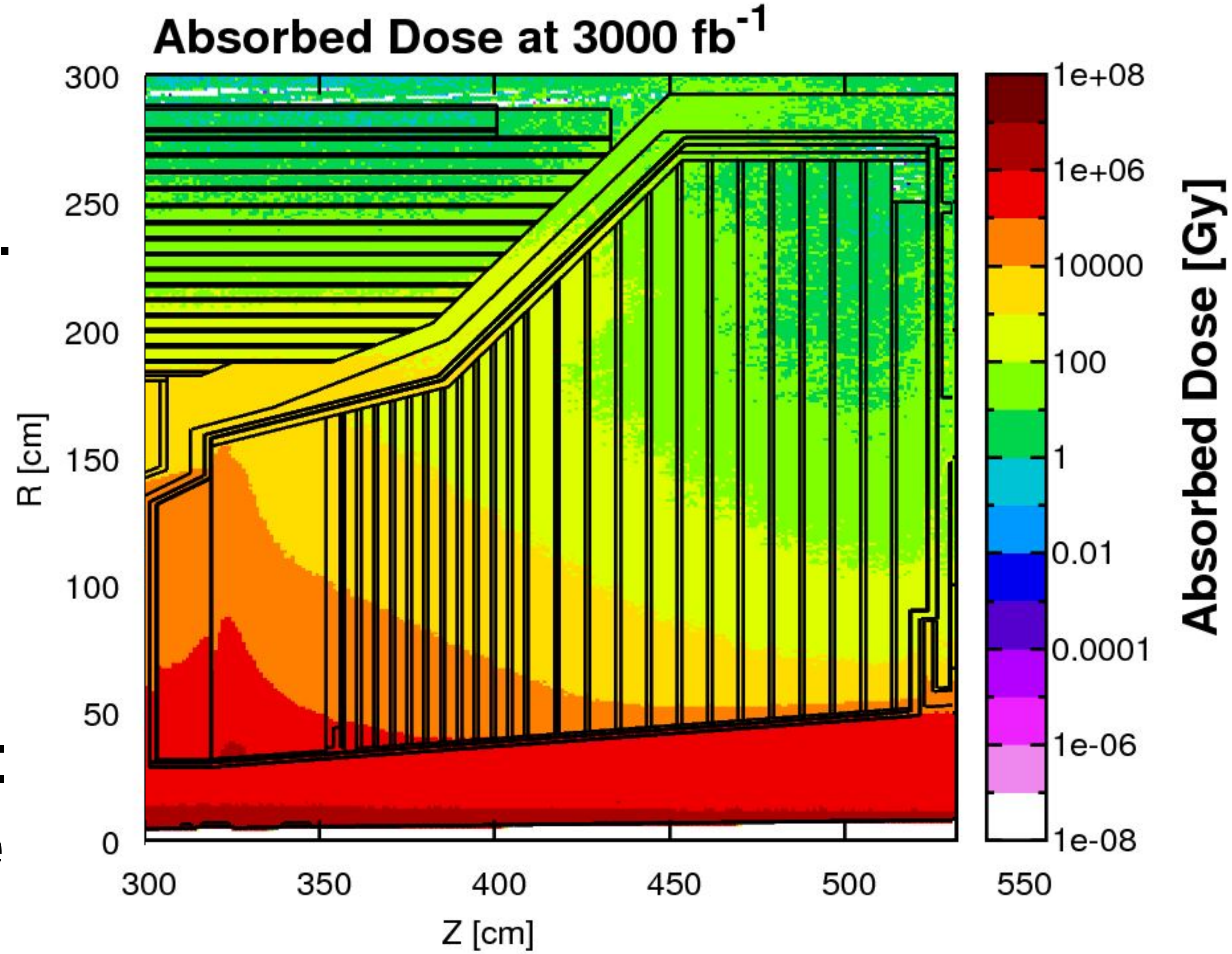
*HGCal is used as a *demonstration* only

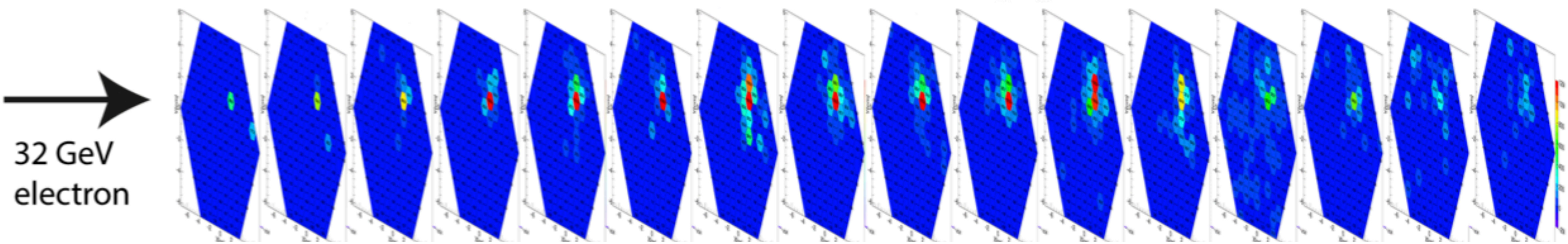
CMS High-granularity Calorimeter

Over 6M channels.
52 layers of Si+absorber.

Data generated @rates up to 380 Gbps/module (40Mhz BX rate)

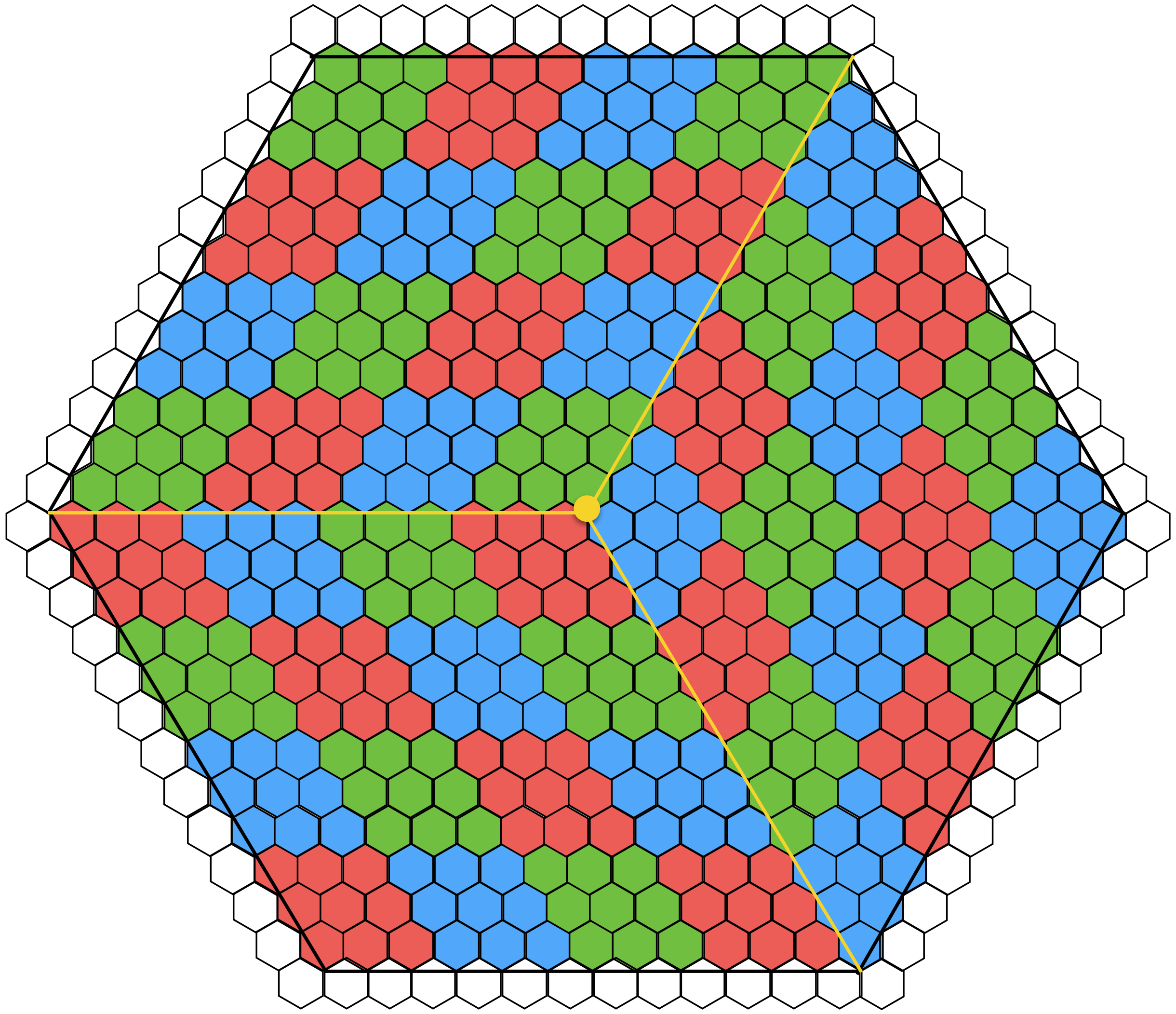
Target trigger bandwidth: 2.56 to 6.4 Gbps/module



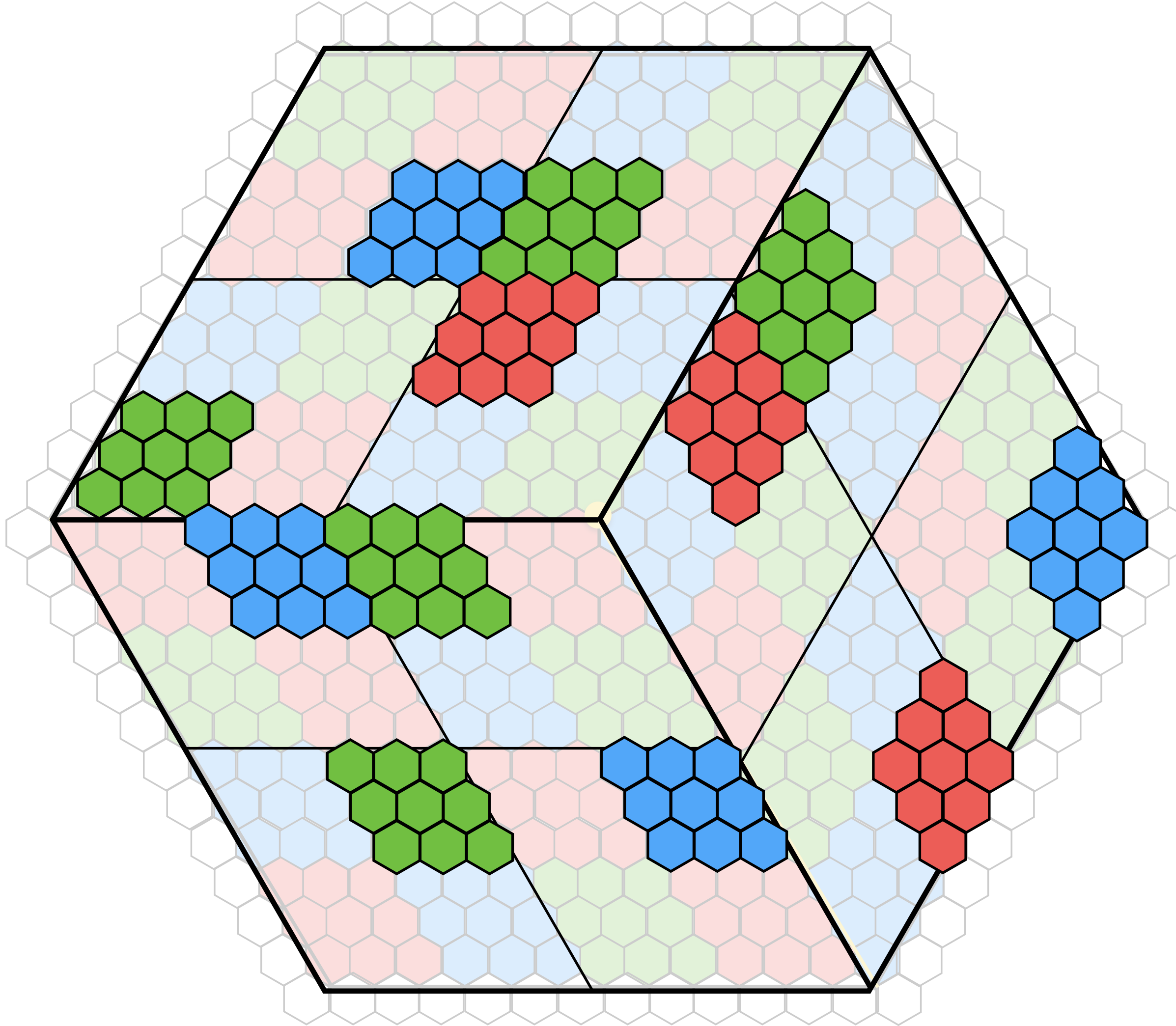


HGCal trigger path

- 40MHz readout requires sending 3x3 "trigger cells" (TCs)
  - Further compression in ECON-T concentrator ASIC



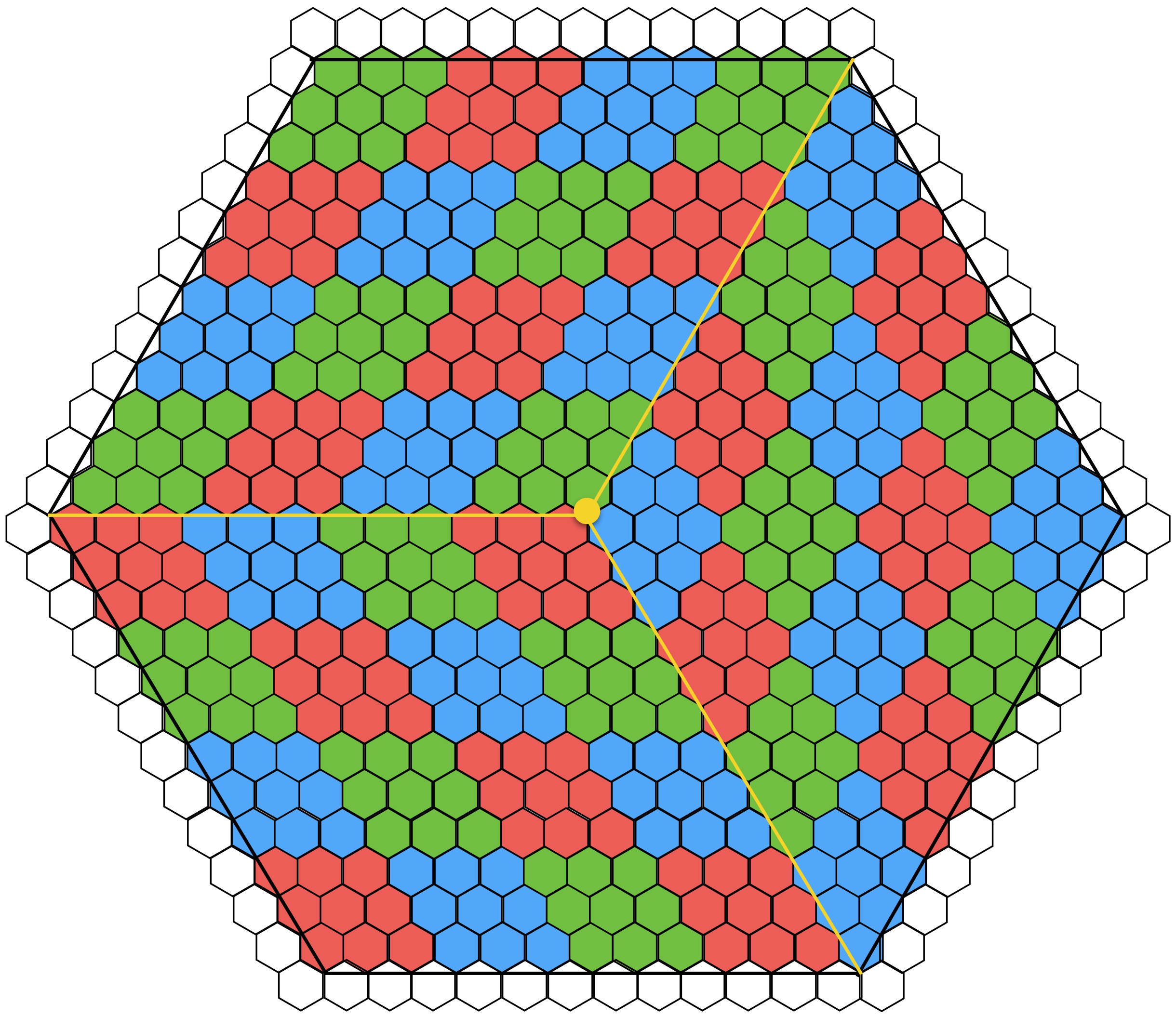
432 Silicon pads
→ 48 TCs / module



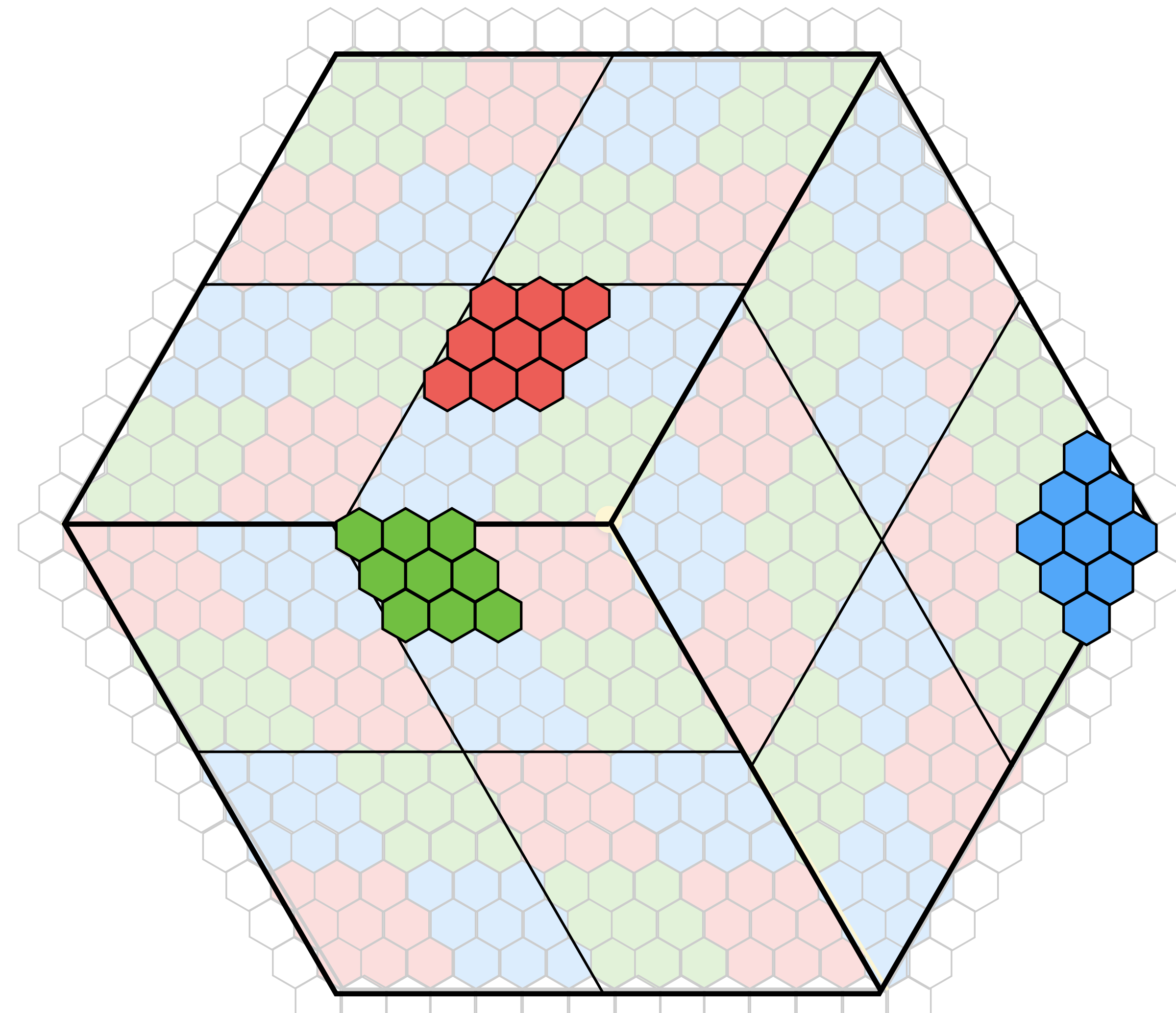
After concentrator ASIC
aggregate 12 sums (high-occ.)

HGCal trigger path

- 40MHz readout requires sending 3x3 "trigger cells" (TCs)
  - Further compression in ECON-T concentrator ASIC



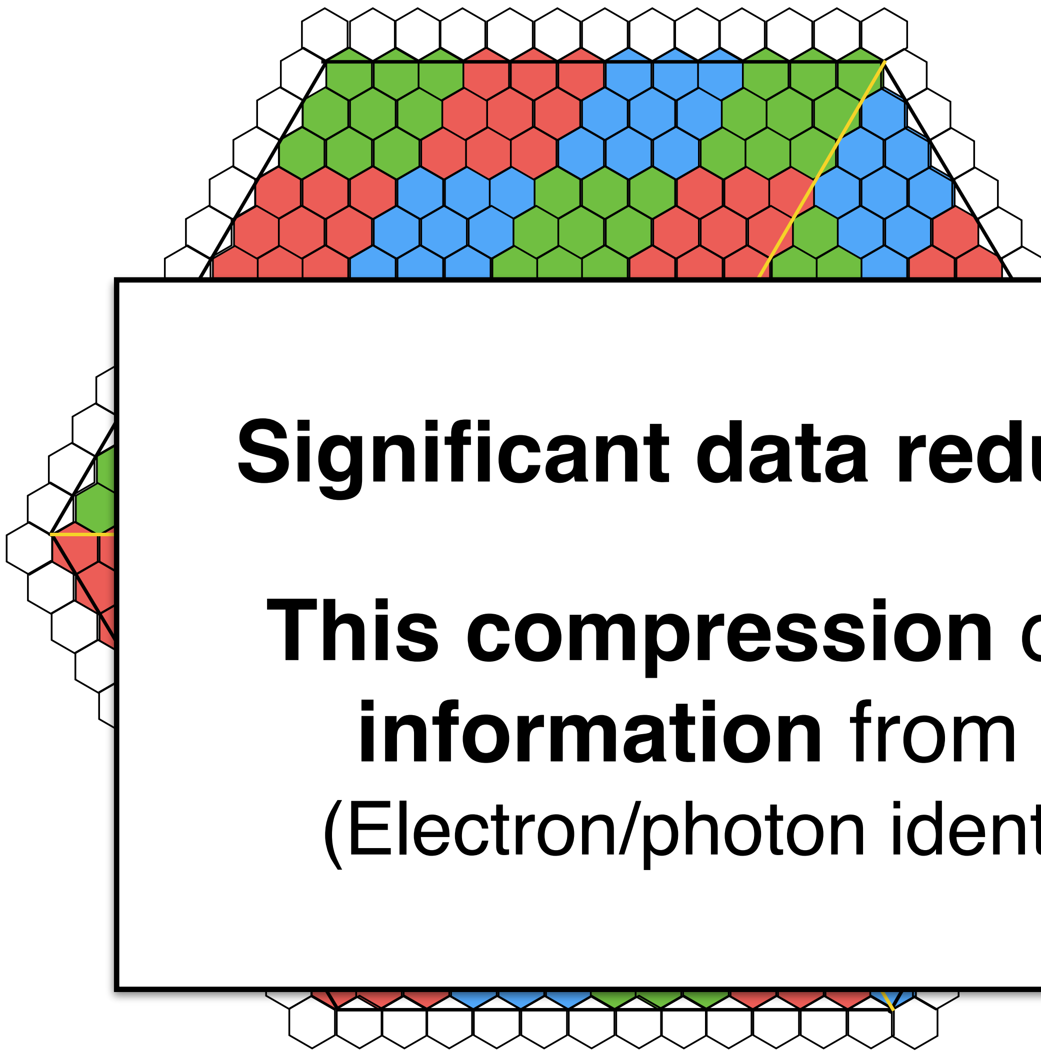
432 Silicon pads
→ 48 TCs / module



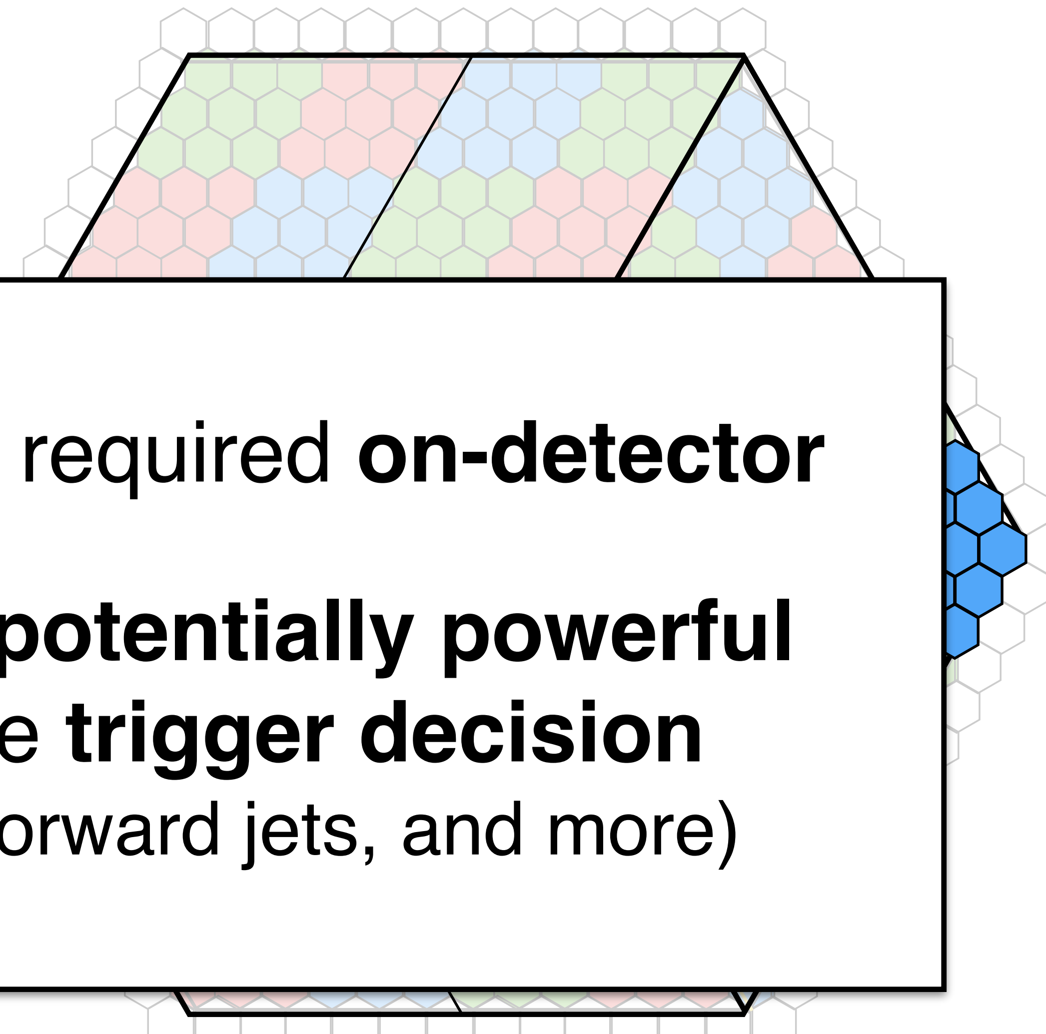
After concentrator ASIC
aggregate 3 sums (low-occ.)

HGCal trigger path

- 40MHz readout requires sending 3x3 "trigger cells" (TCs)
  - Further compression in ECON-T concentrator ASIC



432 Silicon pads
→ 48 TCs / module

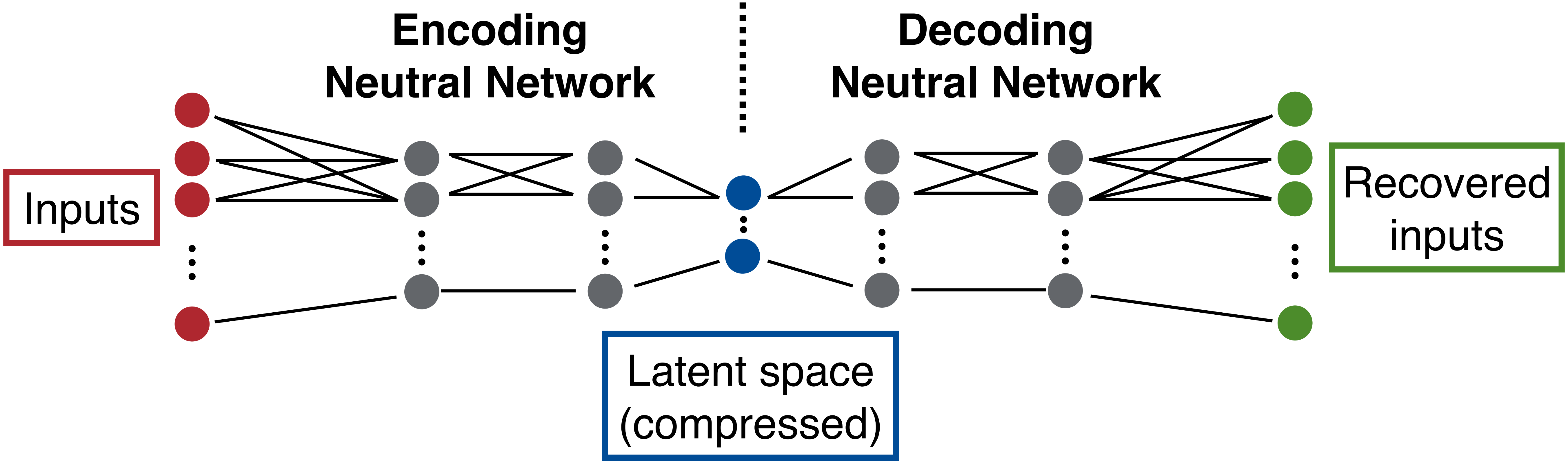


After concentrator ASIC
aggregate 3 sums (low-occ.)

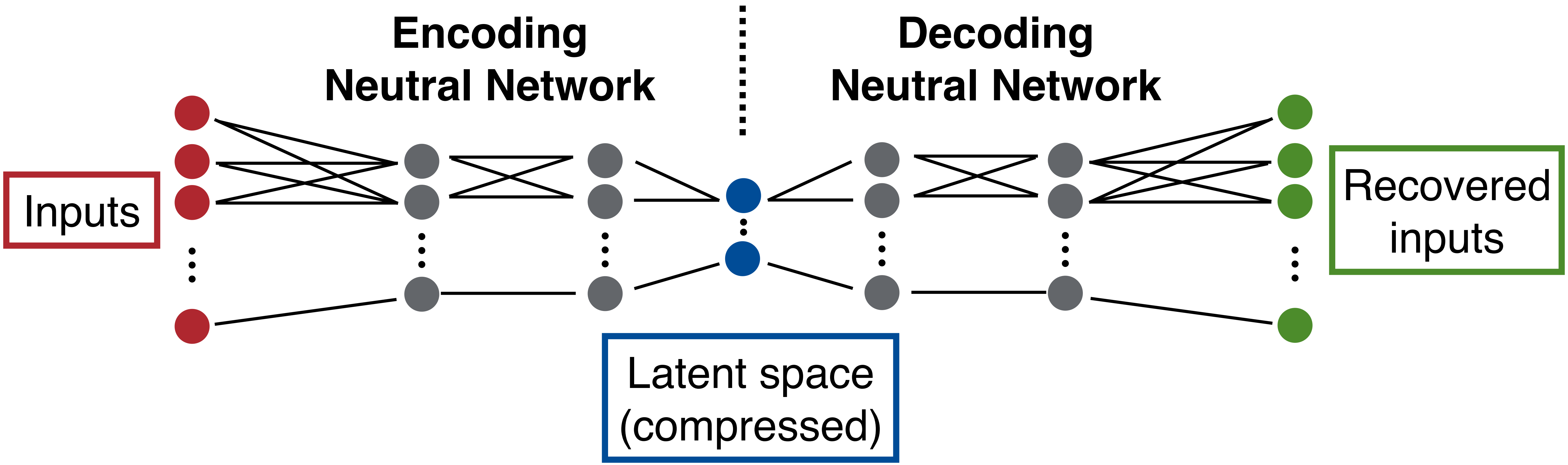
Significant data reduction is required **on-detector**

This compression discards **potentially powerful information** from use in the **trigger decision**
(Electron/photon identification, forward jets, and more)

Auto-encoder concept



Auto-encoder concept



Parameters are fully-reconfigurable!

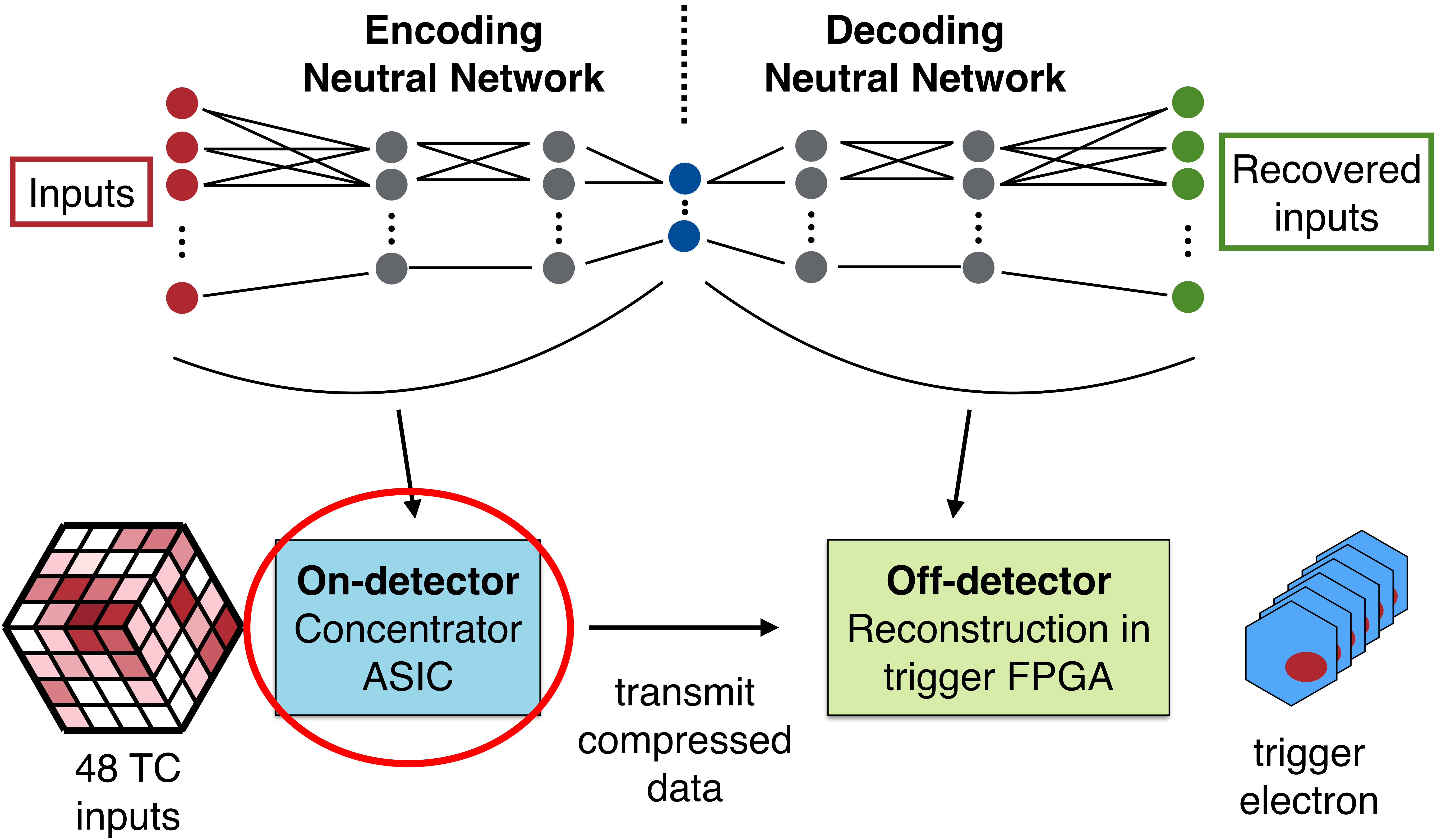
Simplest case:
Fully-connected NN

$$y_i = \sigma(w_{ij} x_i + b_i)$$

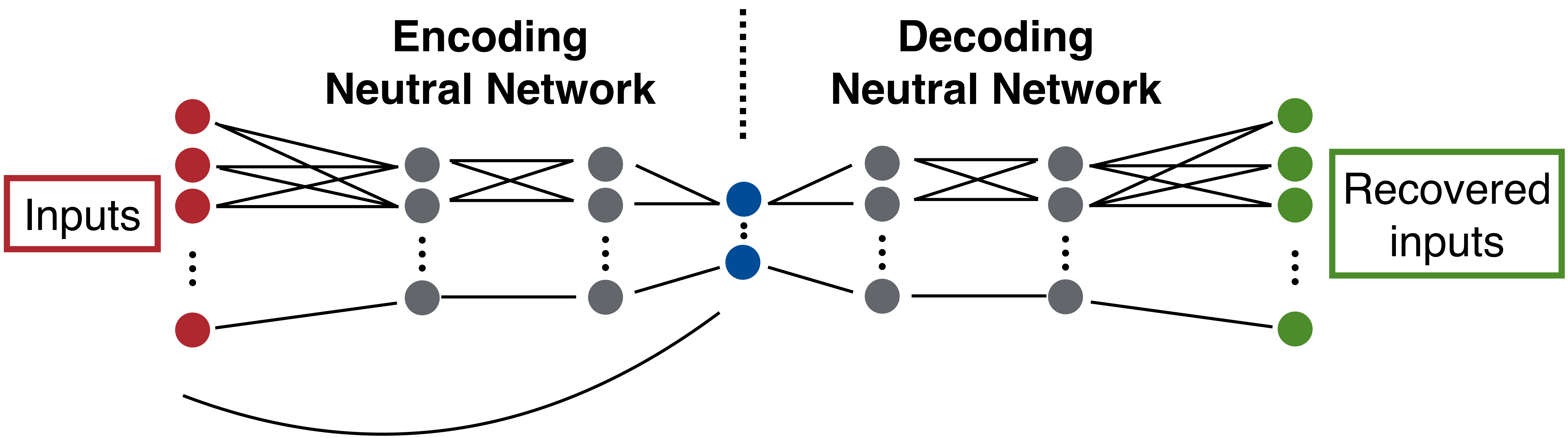
Non-linearity, e.g. $\sigma(x_i)=\max(x_i,0)$

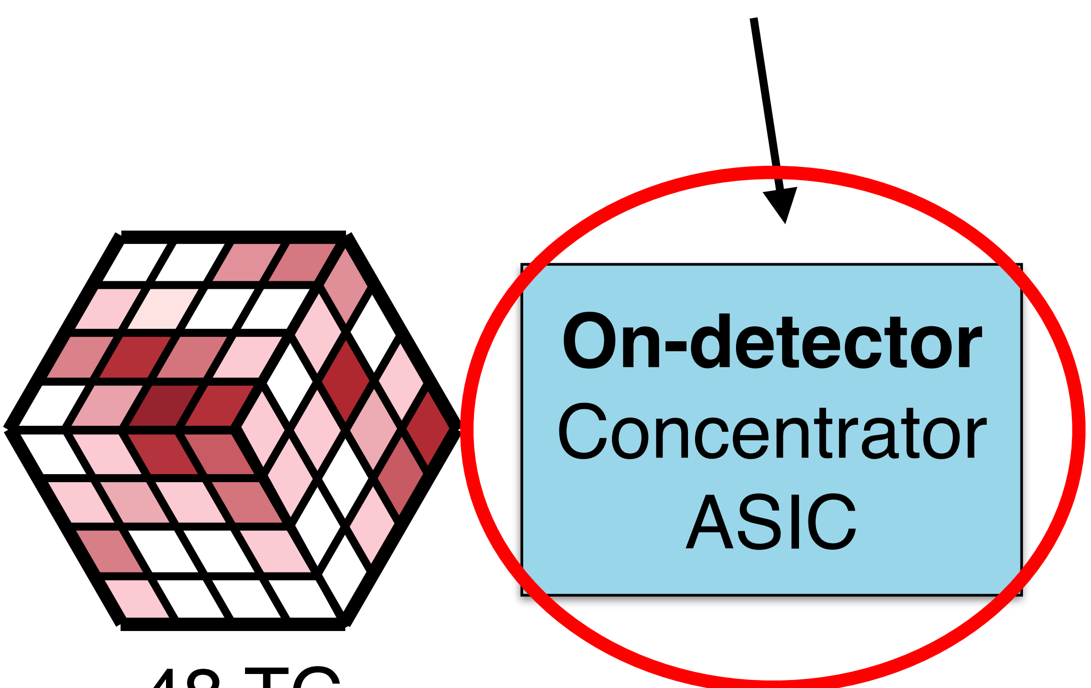
Matrix multiplication

Auto-encoder concept



Auto-encoder concept





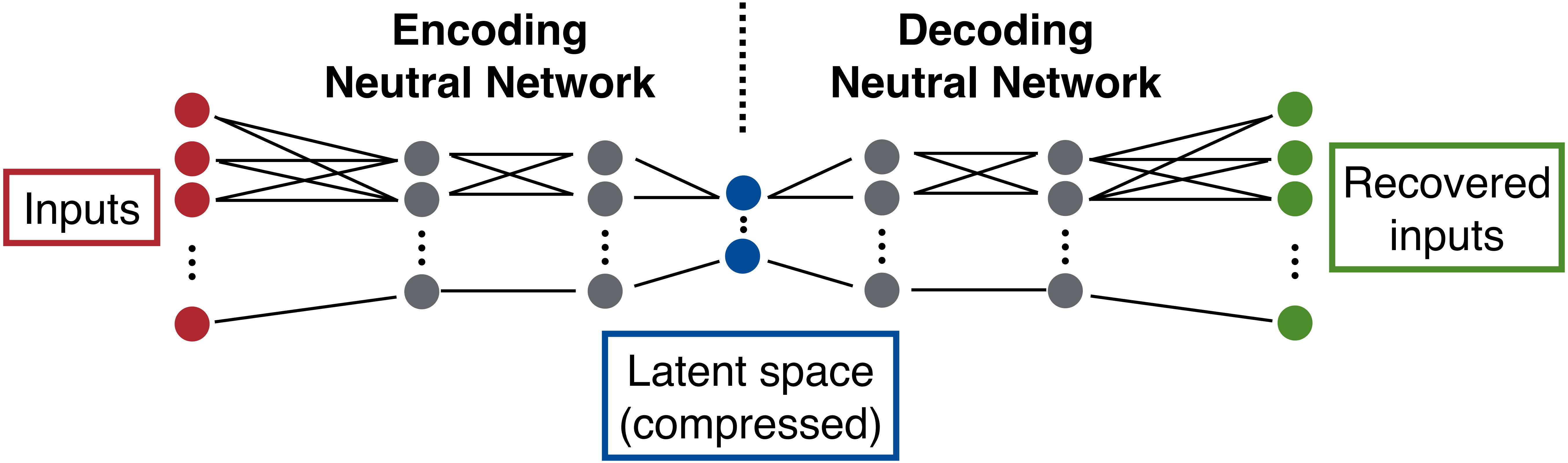
48 TC
inputs

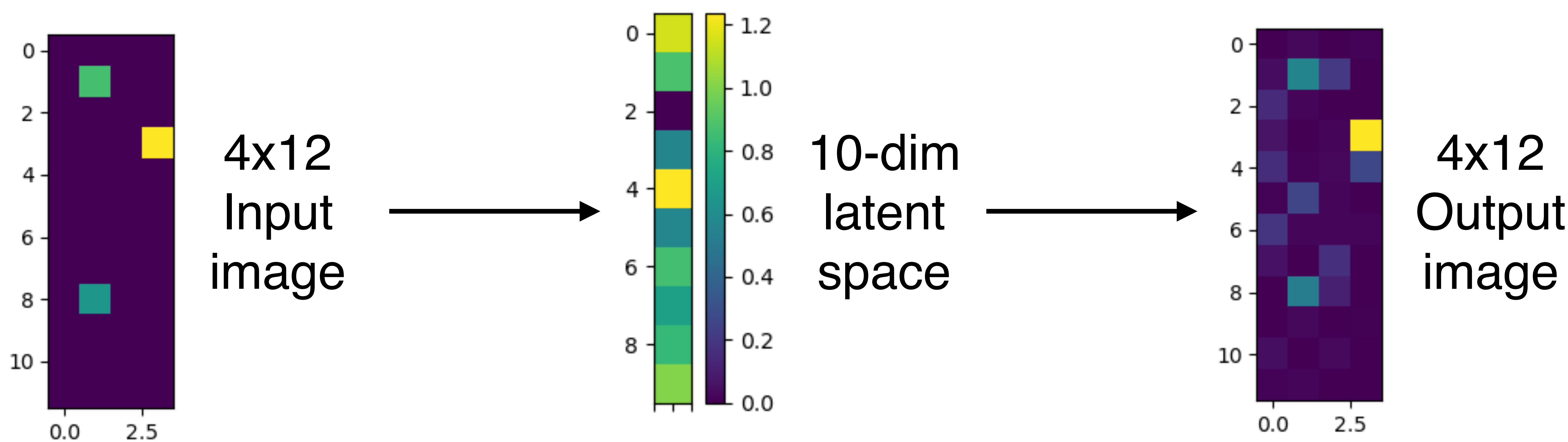
Control **area, power** by minimizing
 # bit operations:

$$N_{\text{ops}} \cdot (b_{\text{inputs}} b_{\text{weights}} + \cdots)$$

Encoder multiply-accumulates

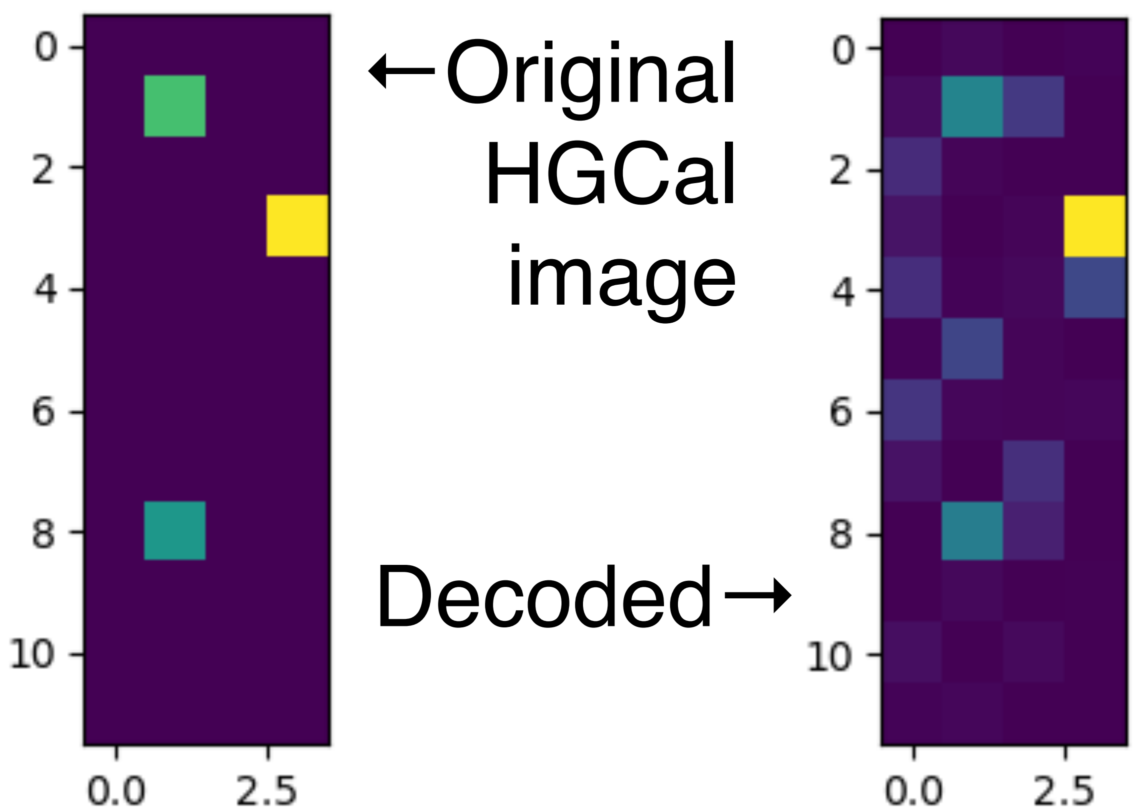
Auto-encoder concept



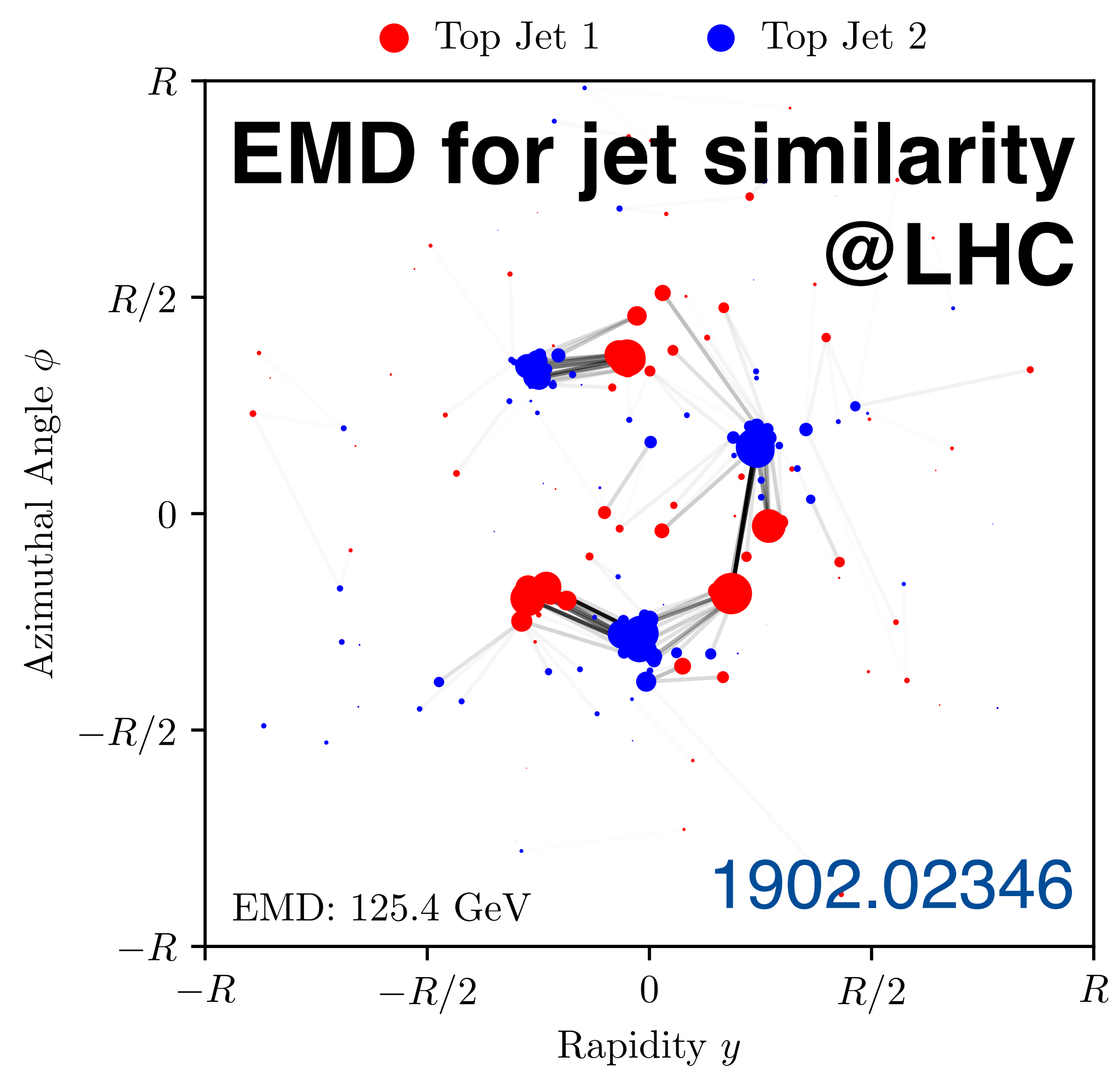


NN inputs and performance metric

- **Data sets:** training+validation samples of jets, electrons, and pileup, using HGCal modules across many layers
- **Image similarity:** Energy mover's distance, measuring the (energy)*(distance) cost of the "optimal transport"

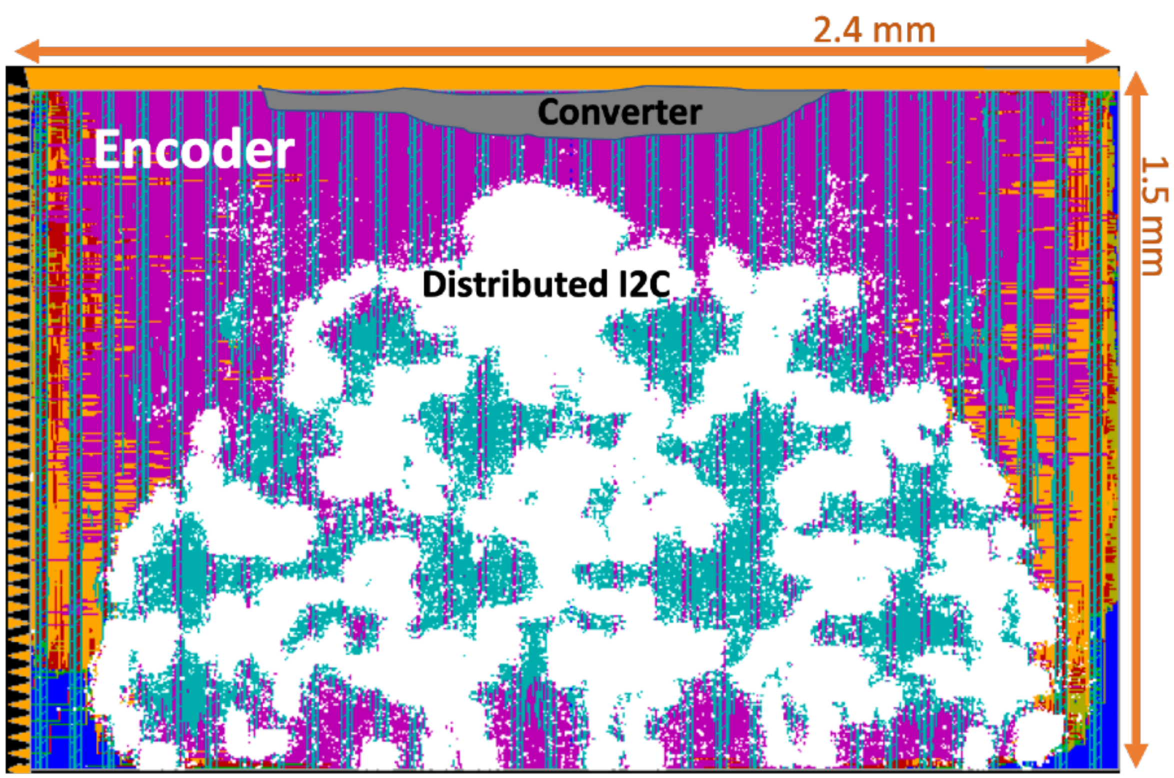


How well did we do?



NN model to ASIC implementation

- **Training:** QKeras enables quantization-aware training
  - "Imaging calorimeter" → Convolutional NN (CNN)
- **Model→RTL:** Translated to HLS via hls4ml, enabling a tight optimization loop combined with CatapultHLS.
  - Hear more on hls4ml in NS-32 from D. Rankin
- **Configurability:** Completely update NN weights via I2C
  - Adapt to changing detector (e.g. radiation effects)
- For full implementation details, see [F. Fahim's talk in NS-24]()

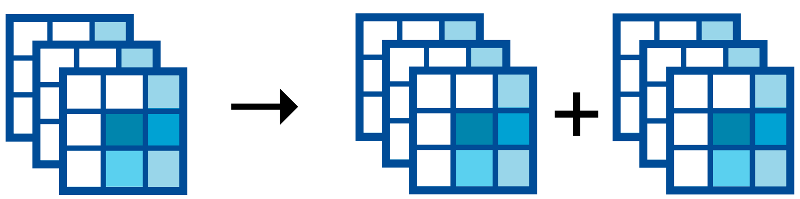


Design floor-plan

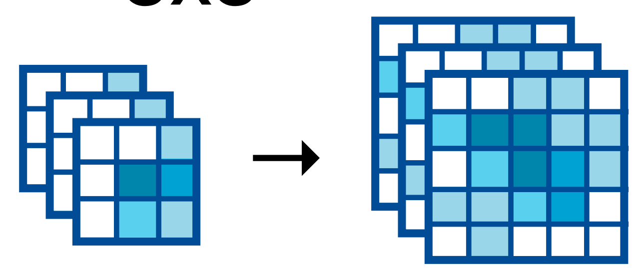
NN architecture exploration

- NN complexity may improve performance at the cost of a larger, more power-hungry design
  - Begin with **simple CNN**: 1 convolution + 1 dense layer
- Many possible variations were investigated:

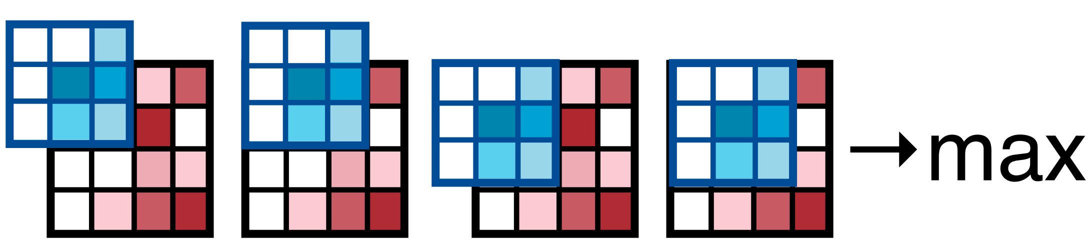
Extra layers:
e.g. 2 convolutions



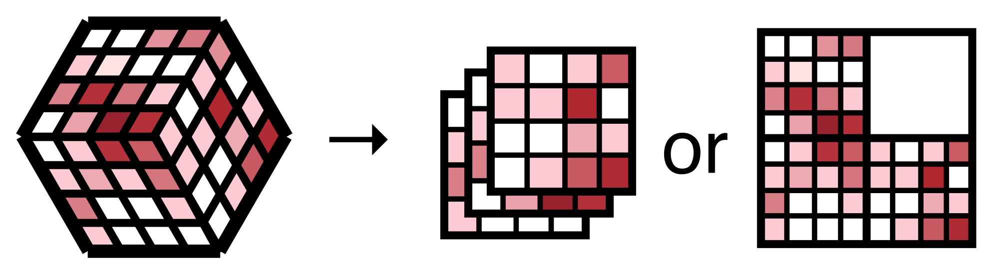
Larger kernel size:
3x3 → 5x5



'Pooling' convolution outputs:



2d/3d convolution inputs:

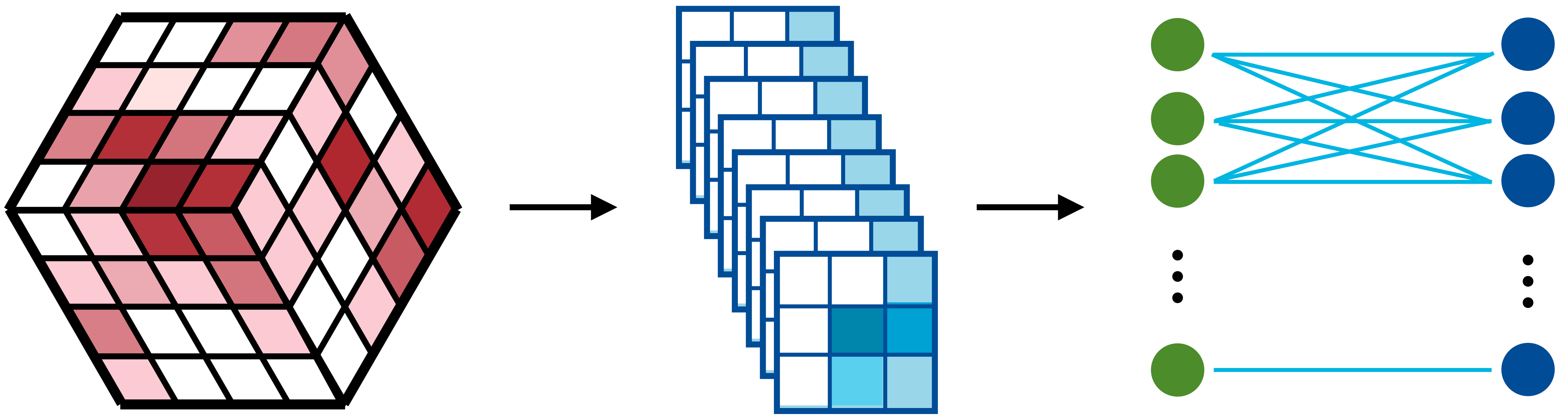


NN architecture exploration

- NN complexity may improve performance at the cost of a larger, more power-hungry design
  - Begin with **simple CNN**: 1 convolution + 1 dense layer
- Many possible variations were investigated:

Model configuration	EMD	Parameters	Ops/eval
Nominal	2.06	2288	11152
Extra conv layer	2.14	+26%	+333%
Extra dense layer	2.00	+12%	+5%
5x5 kernel	1.86	+17%	+110%
2x2 pooling	1.57	-67%	-26%
2d inputs	1.47	+173%	+76%
Non-NN "Aggregation" in 2x2 / 4x4 sums	4.07 / 4.77	n/a	n/a

Optimizing bit-wise precisions (I)



48 inputs (trigger cells)

8 Conv filters

128 features

16 **outputs**

Input size reduced by normalizing to sensor maximum (22→8 bits).

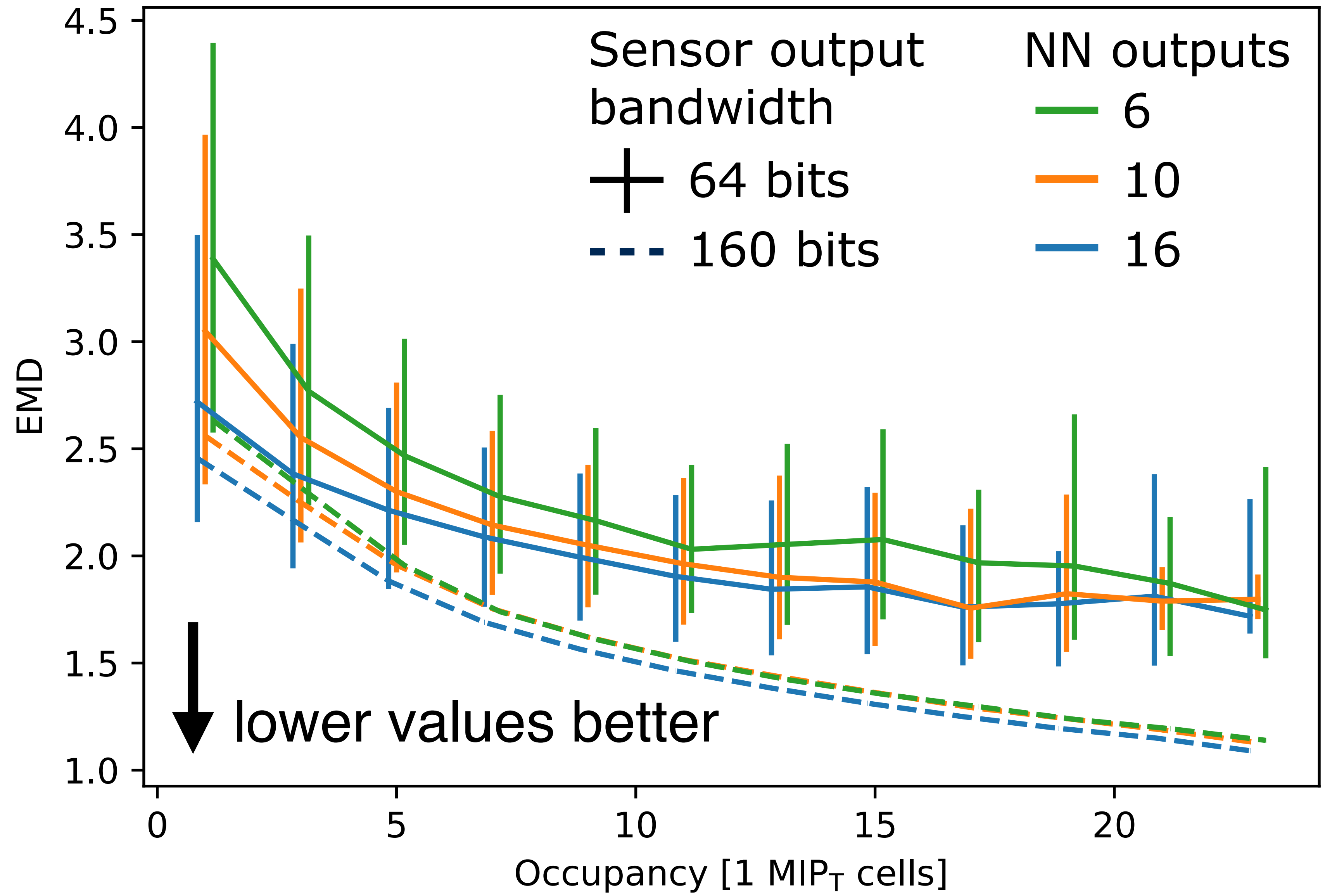
Precision of all **internal weights** may be optimized (within ~4mm² ASIC area constraint).

Output precision is set by occupancy. Algo is configurable from 48 to 144 bits.

Optimizing bit-wise precisions (II)

- Better to perform many low-precision calculations or fewer with higher precision?
  - Find optimal weight precision, while keeping area fixed.

- Better encoding using **many low-precision calcs.**
- True for high and low-bandwidth scenarios.



Conclusions

- We have presented a NN encoder targeting the CMS HGCal concentrator ASIC
    - Our design profits from a tight optimization loop using quantized training and HLS allows for rapid iteration.
- Expanded on-detector processing may enhance the physics performance of off-detector trigger logic
    - Suggests means to exploit fine granularity in the trigger
- Reconfigurability is key to adapt to changing conditions and benefit from future model improvements.
- Beyond the HGCal, the design flow and optimization tools explored here might extend to Intelligent Detectors in data-rich environments across HEP experiment.

Thanks to all Collaborators!

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

L. Carloni, G. Di Guglielmo

BROWN

M. Kwok

Fermilab

F. Fahim*, B. Hawks, CH,
J. Hirschauer, N. Tran*

FLORIDA TECH

D. Noonan

Northwestern

Y. Luo, S. Ogrenci-Memik

Thanks also to the FNAL ECON team, CMS HGCal and Fast Machine Learning communities!

*also Northwestern