# Colony of NPUs: Scaling the Efficiency of Neural Accelerators

**Babak Zamirai**, Daya S Khudia, Mehrzad Samadi, and Scott Mahlke

University of Michigan, Ann Arbor

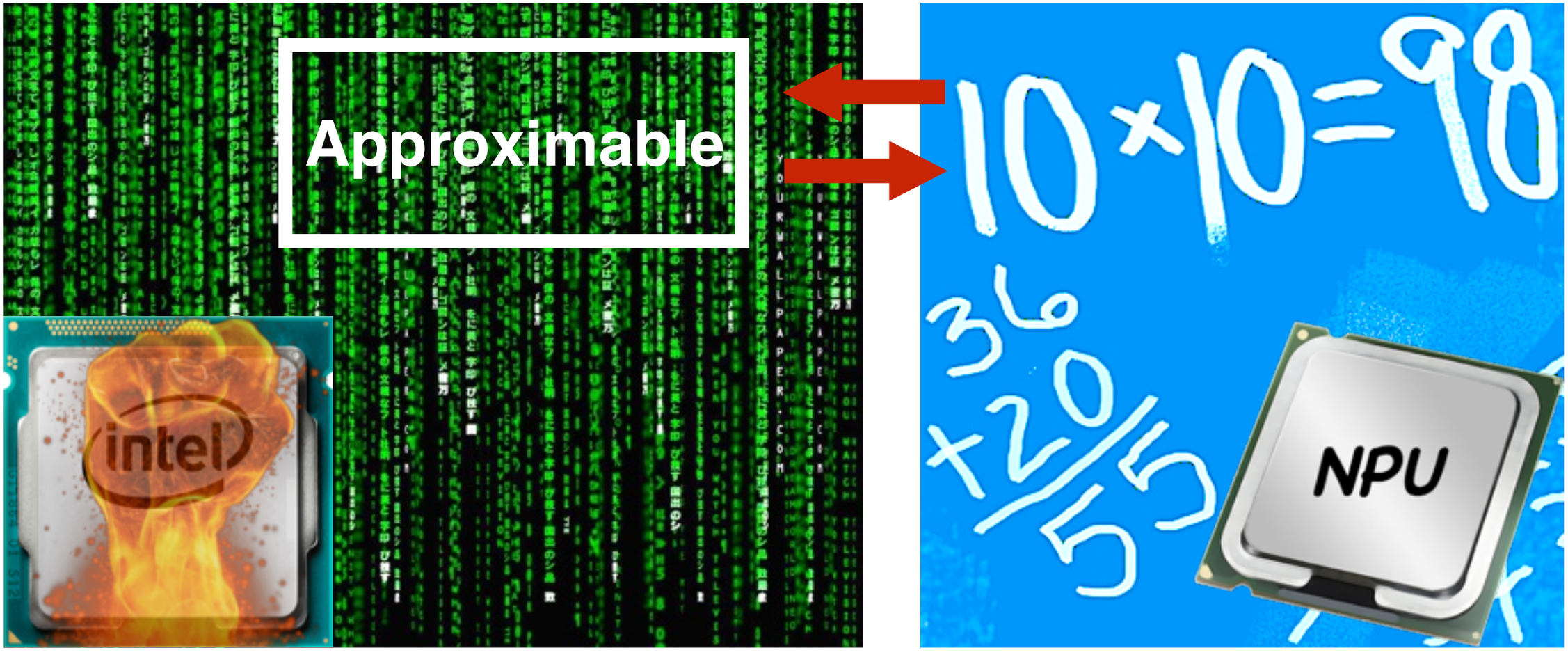
June 2015



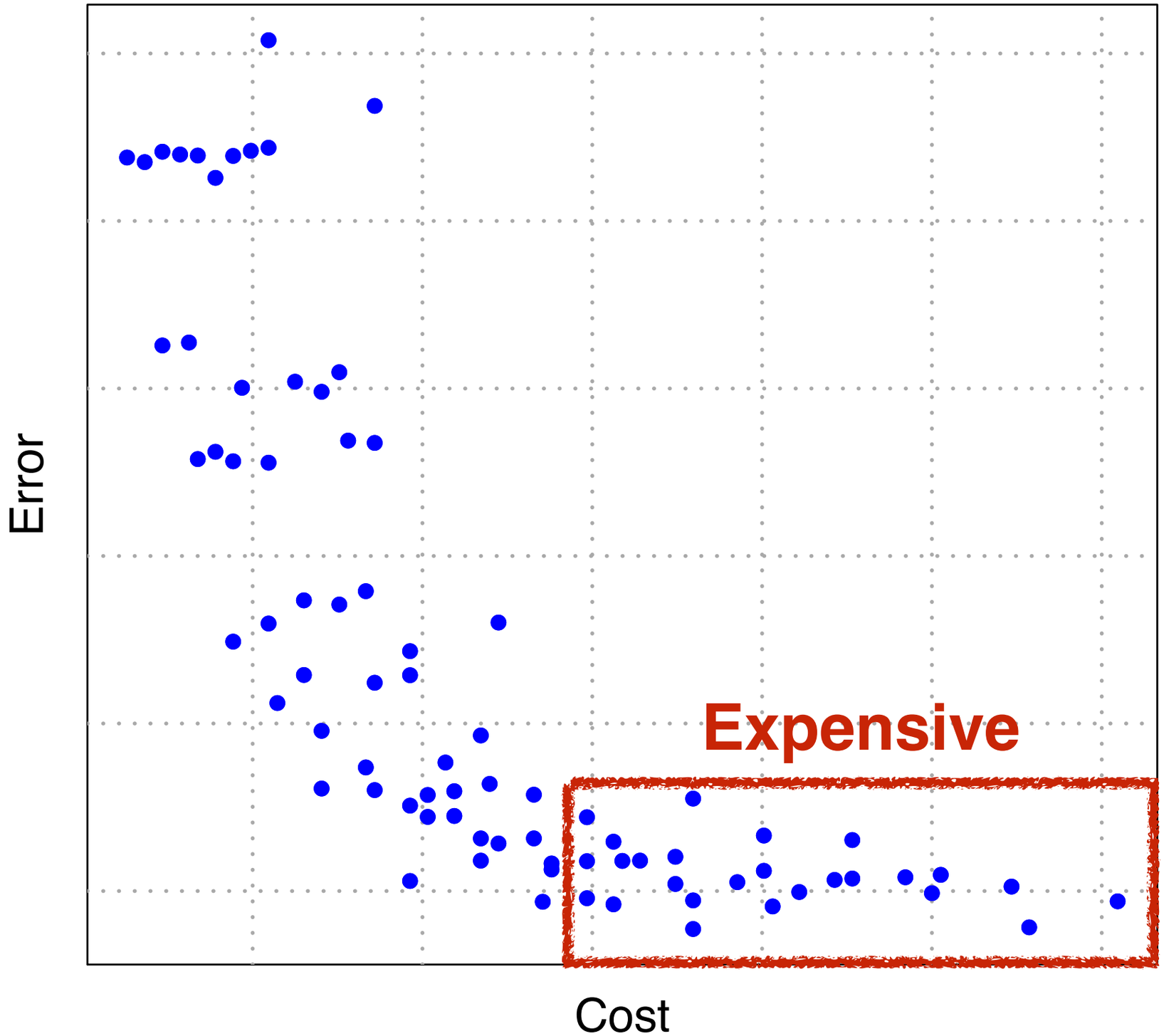
# Neural Processing Unit (NPU)

- **Trade accuracy for**
  - **Performance**
  - **Energy consumption**

**[Esmaeilzadeh et. al., Micro, 2012]**

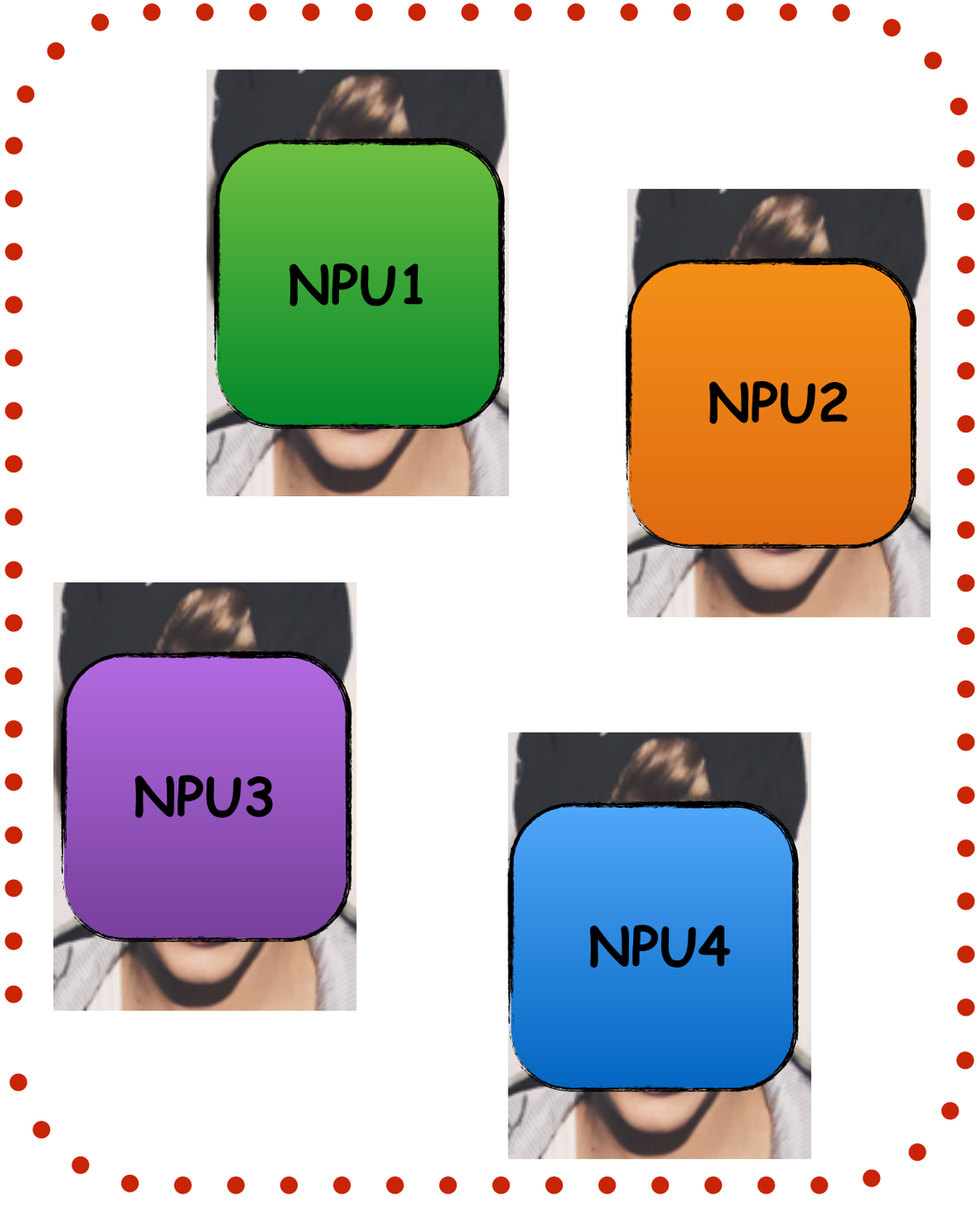


# Better Accuracy is Expensive

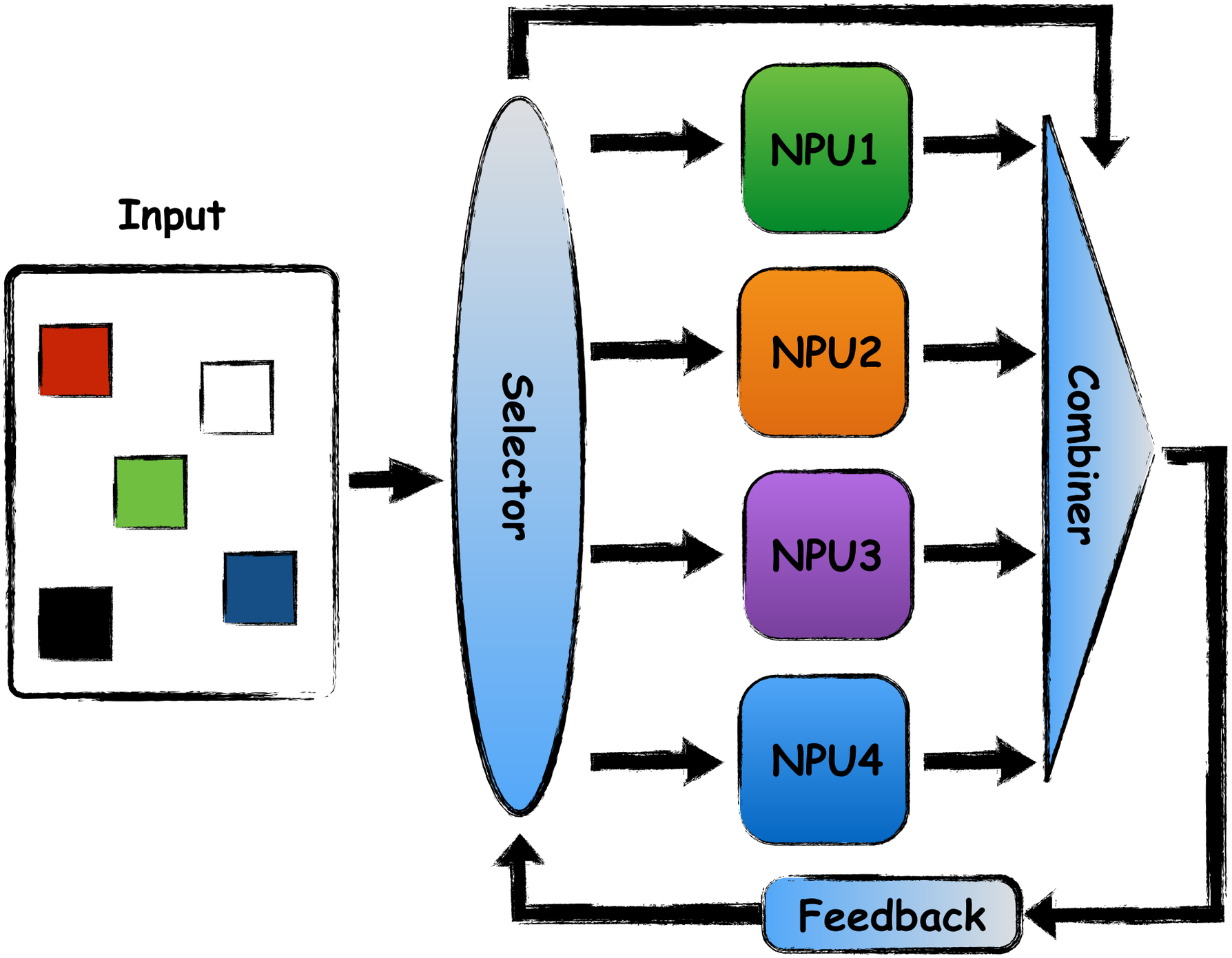


# Boosting Algorithms

- **A set of weak learners create a strong learner**

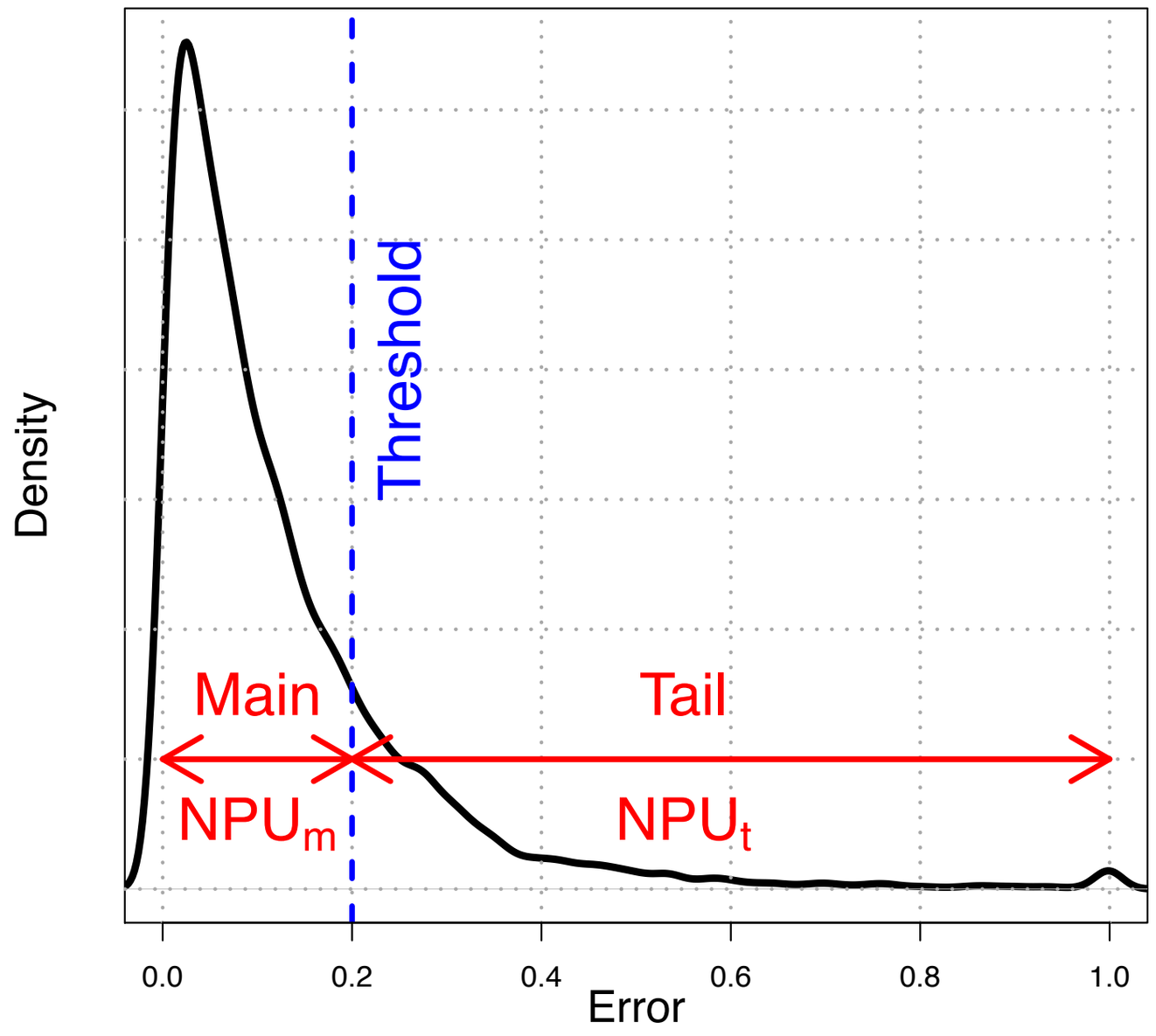


# Colony of NPUs (cNPU)



# Error Distribution

- **Split**
  - **Main part**
  - **Tail part**



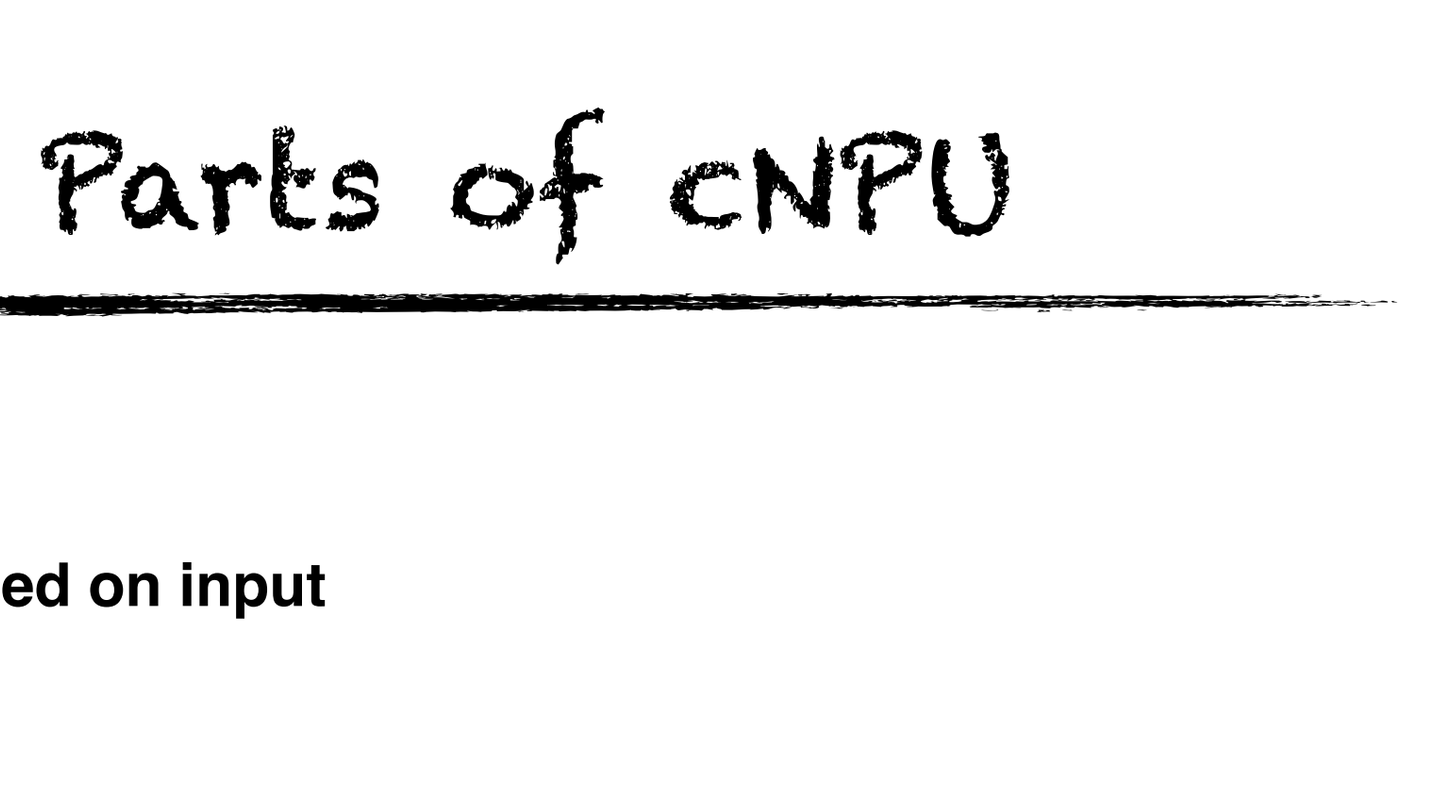
# Main Parts of cNPU

## ✦ Selector

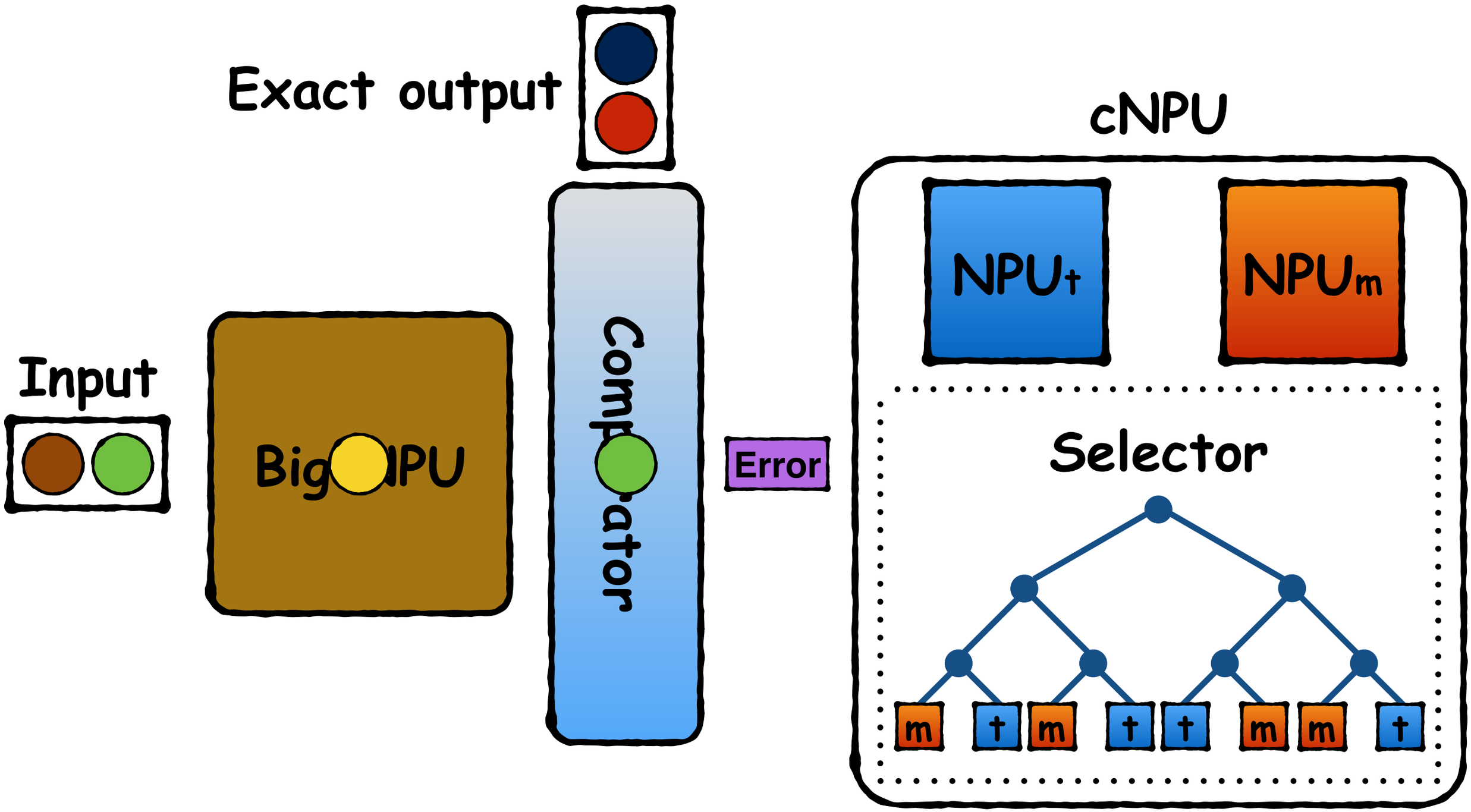
- Categorize error based on input
- Decision tree
- Tiny neural network

## ✦ Combiner

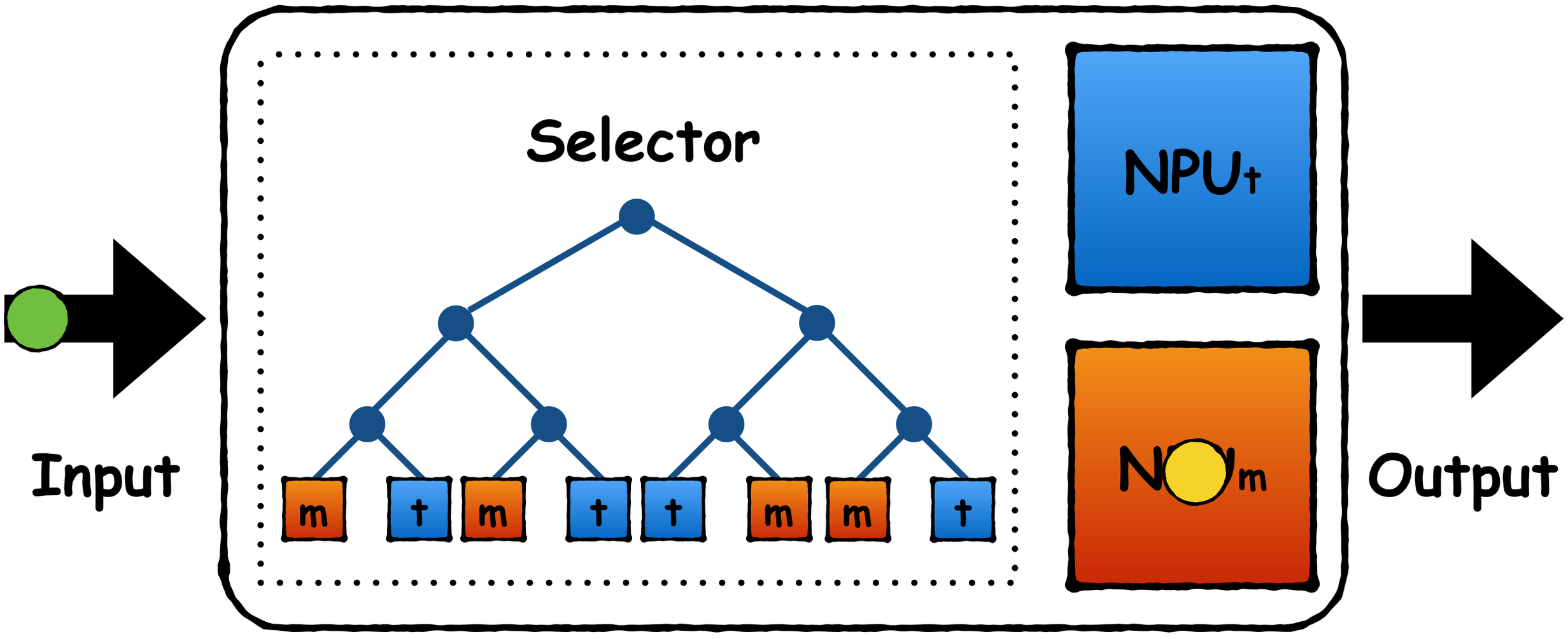
- Trivial



# Training cNPU



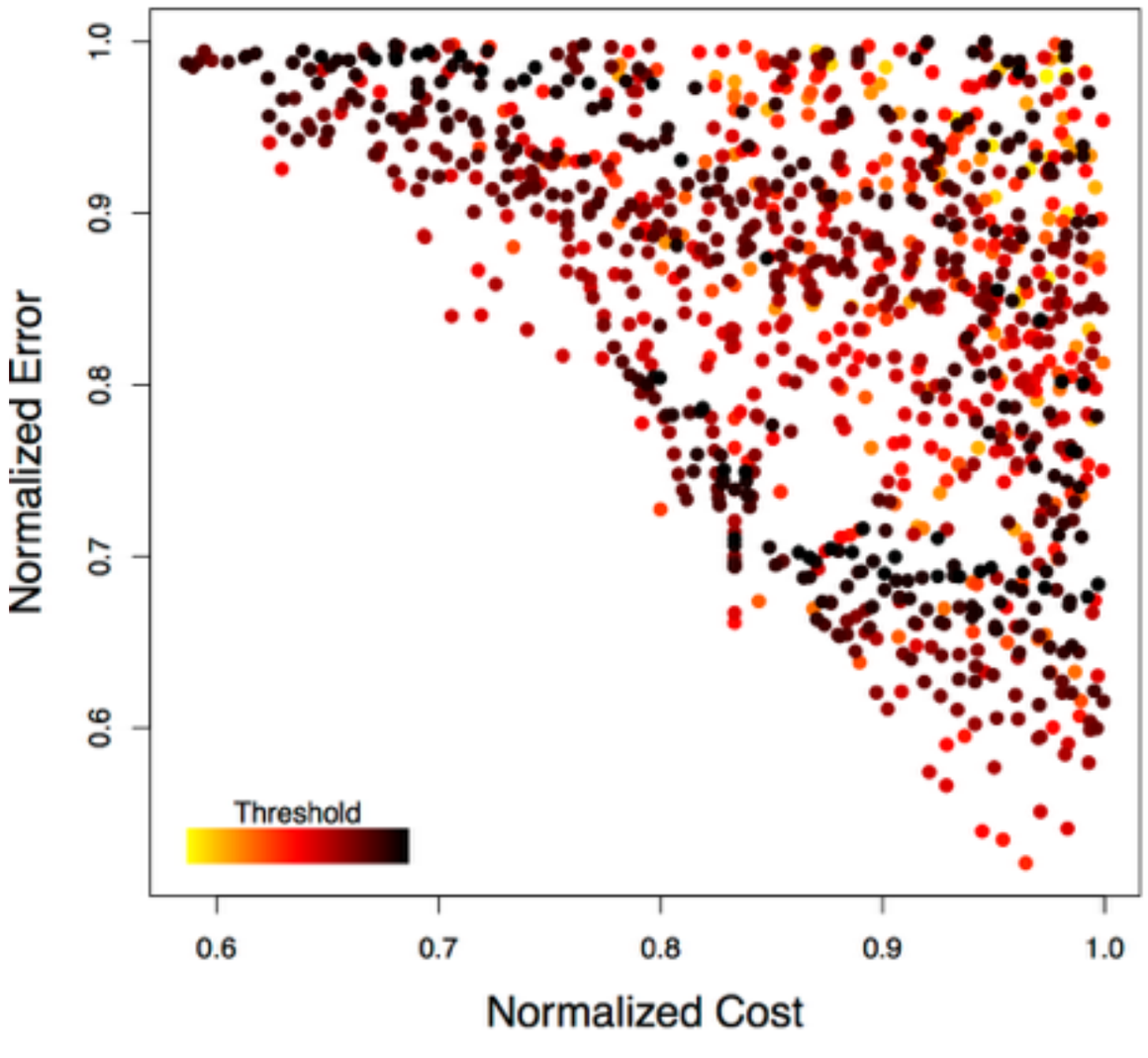
# Runtime



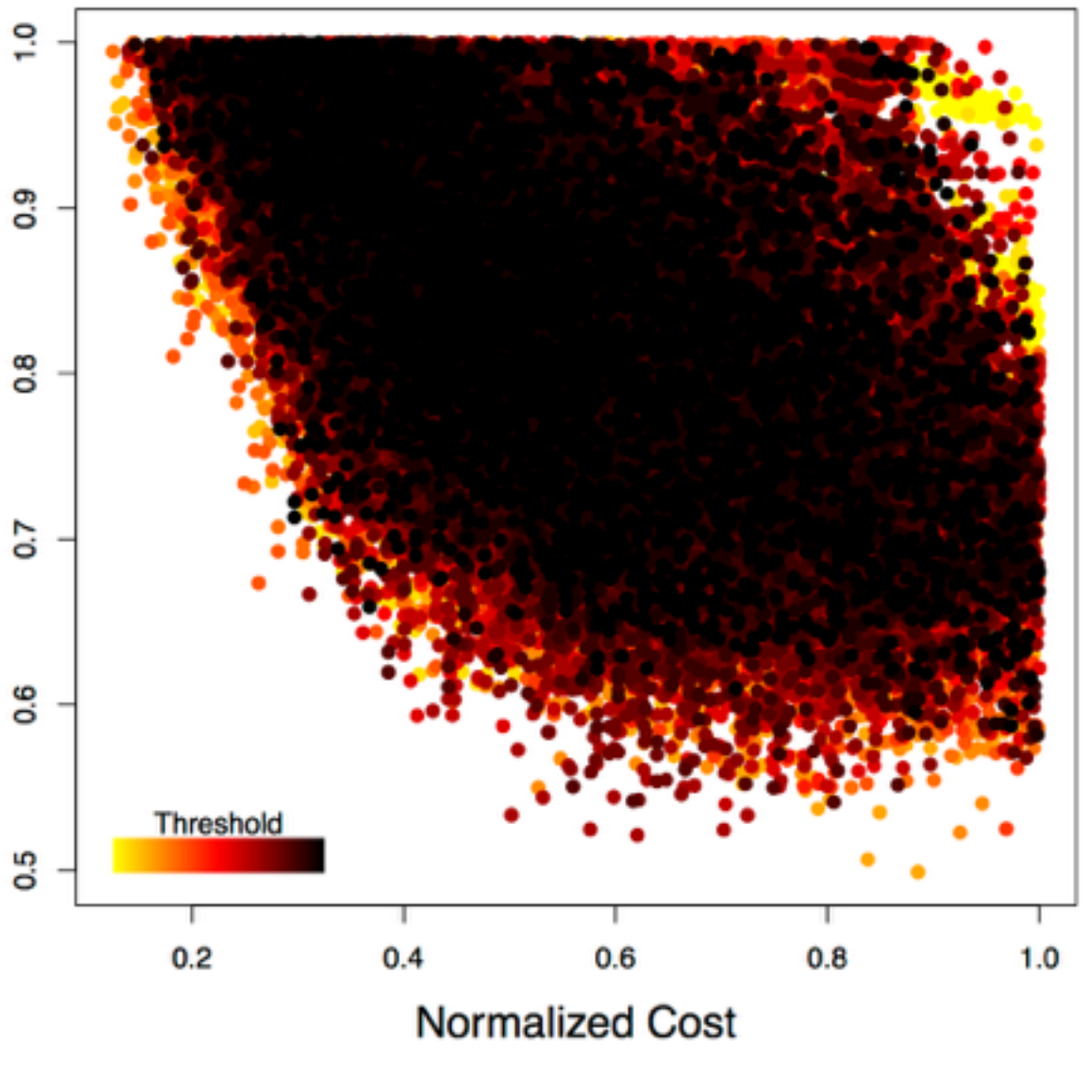
# Opportunity of Improvement

- **Different configurations for $NPU_m$ and $NPU_t$**
- **Different splitting thresholds**

## K-Means (2,2)



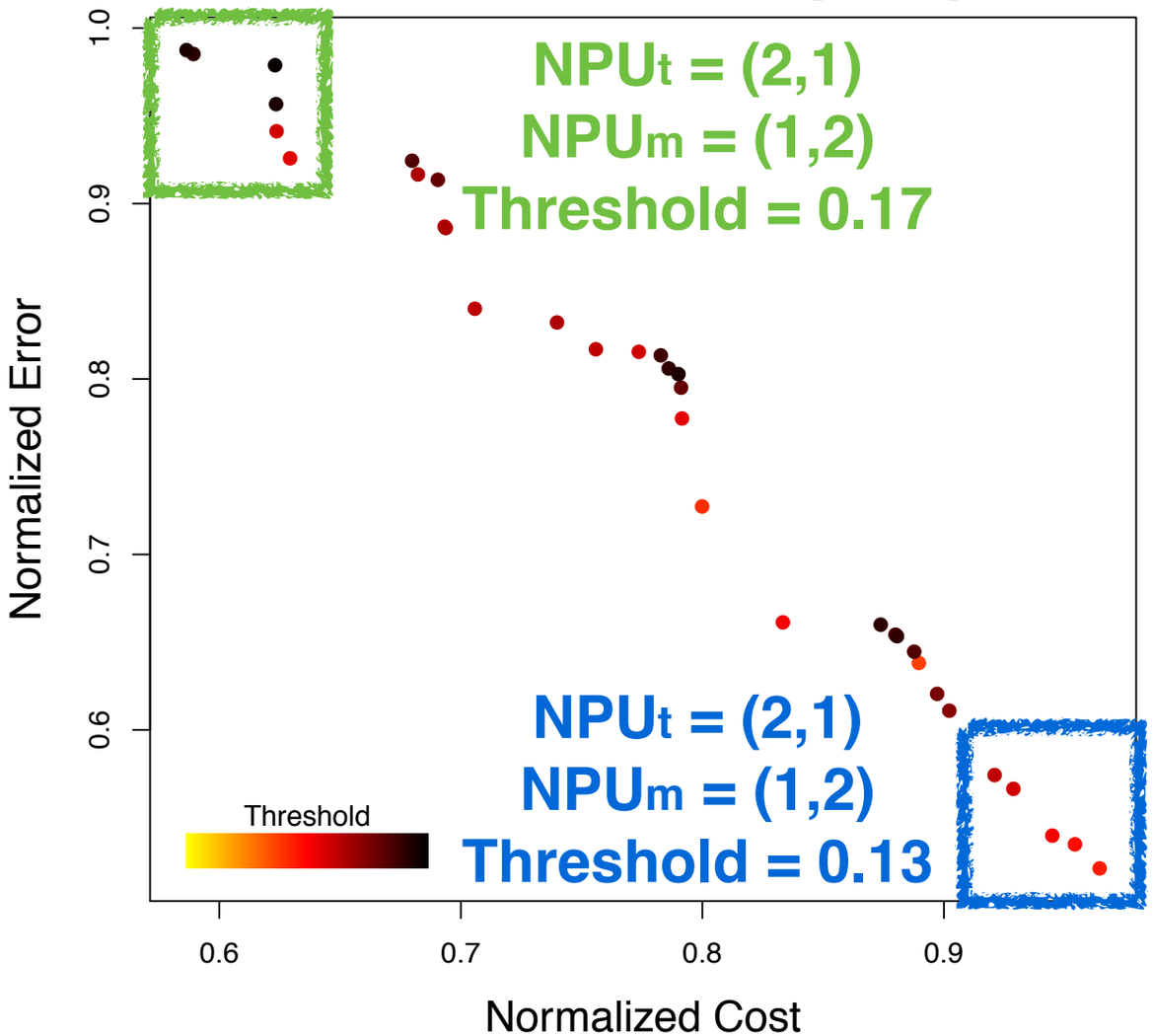

## Black-Scholes (8,8)



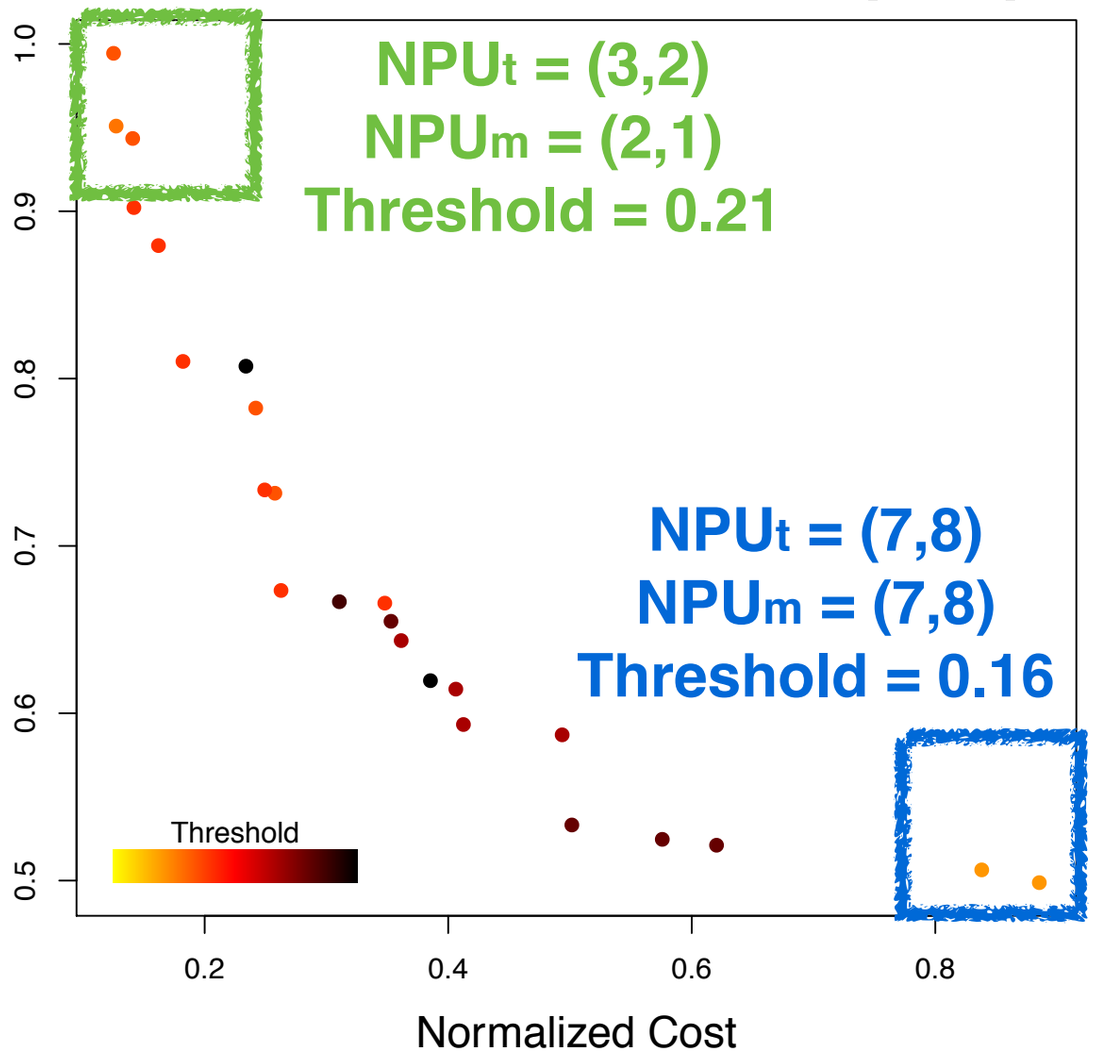
# cNPU Configuration

- **Reduce cost**
- **Improve accuracy**

## K-Means (2,2)

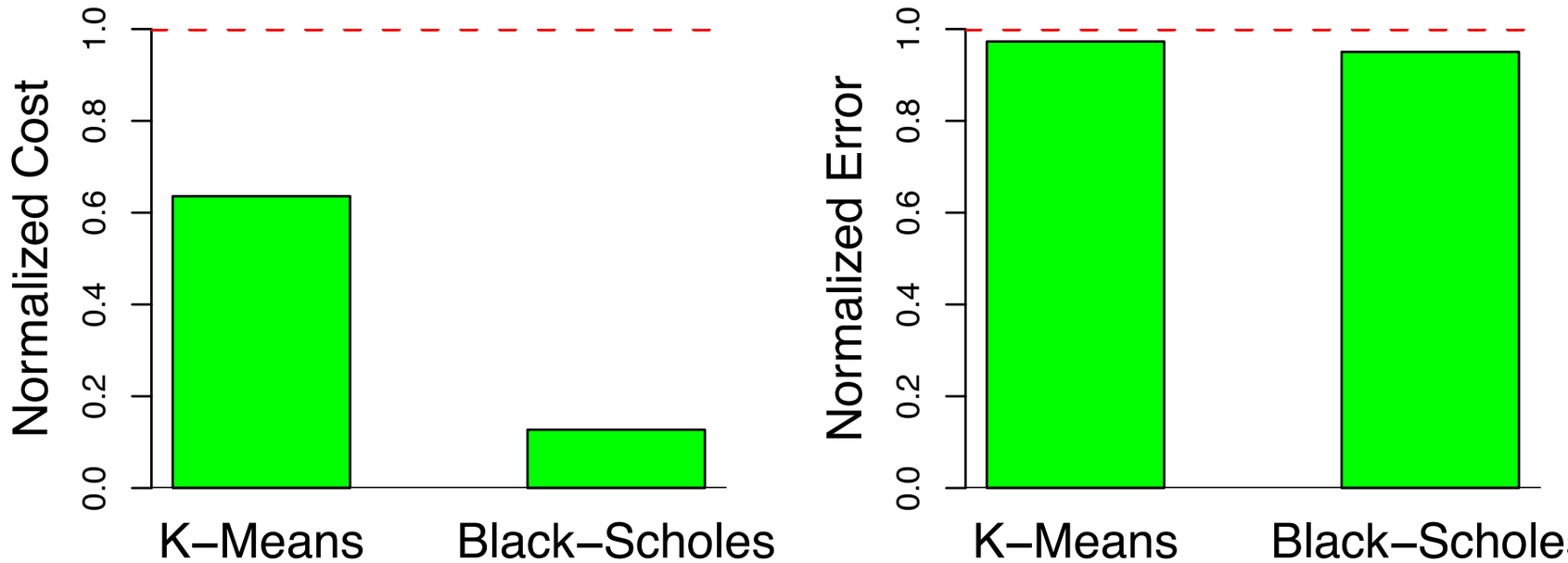



## Black-Scholes (8,8)



# Evaluation

**Minimizing Cost**



# Conclusion

- **Better accuracy need bigger NPUs**
- **cNPU is combination of small NPUs**
- **Up to 87% cost reduction (same accuracy)**
- **Up to 1.95x accuracy improvement (same cost)**