ANDANTE

AI FOR NEW DEVICES AND TECHNOLOGIES AT THE EDGE

KPI-aware Optimization and Design

Maen Mallah (Fraunhofer IIS), Ferdinand Pscheidl (Fraunhofer EMFT)

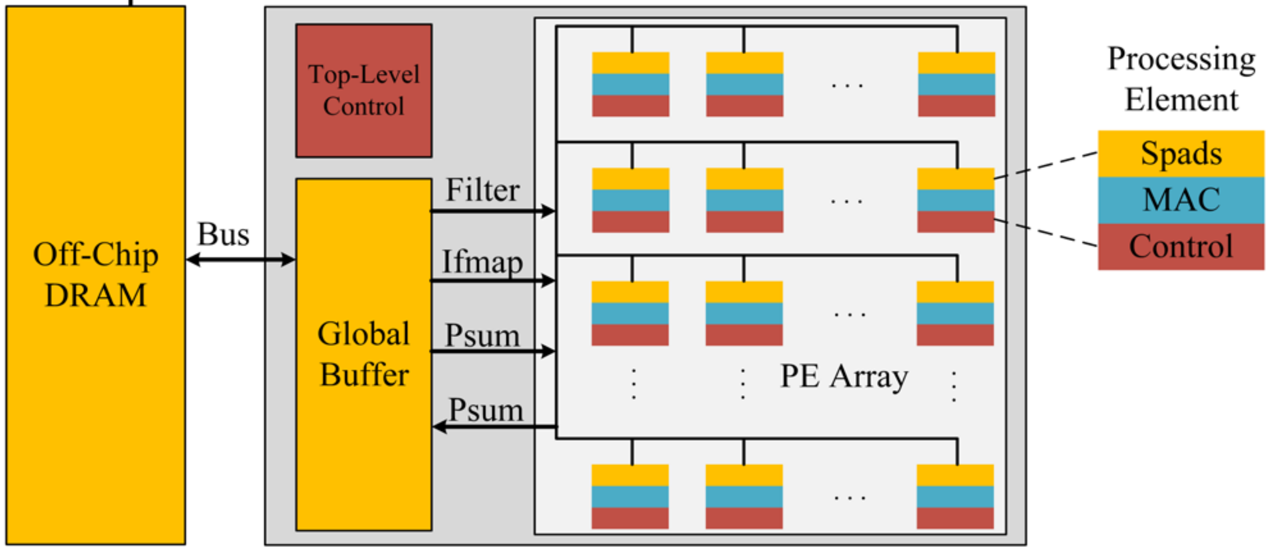
OUTLINE

- ✓ Motivation
- ✓ Proposed training flow
- ✓ Example for KPI evaluation
- ✓ Constraints of training flow
- ✓ Advantages and applications
- ✓ Required models of accelerators
- ✓ Conclusion

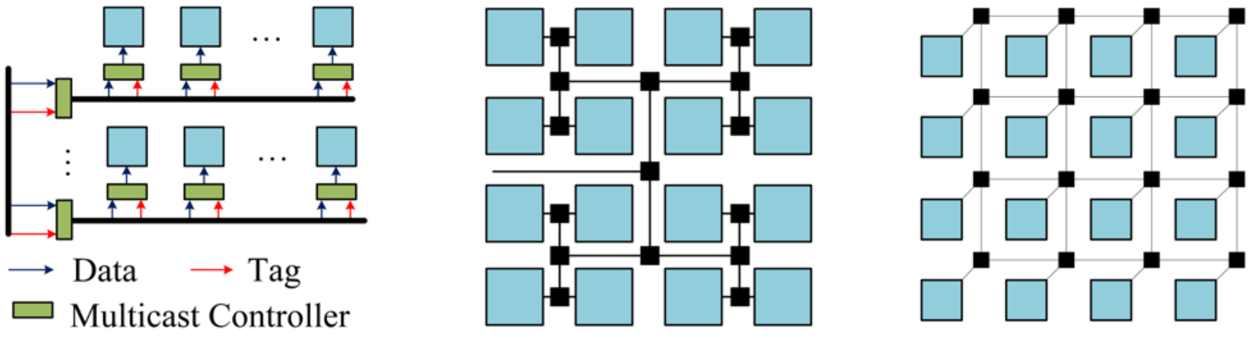
MOTIVATION

- Application-specific ANN/SNN accelerators
- Theoretical TOPS/W alone not meaningful
- Benchmarking based on ANN/SNN models not fair
- **Proposal:** Benchmark based on use-cases (applications)

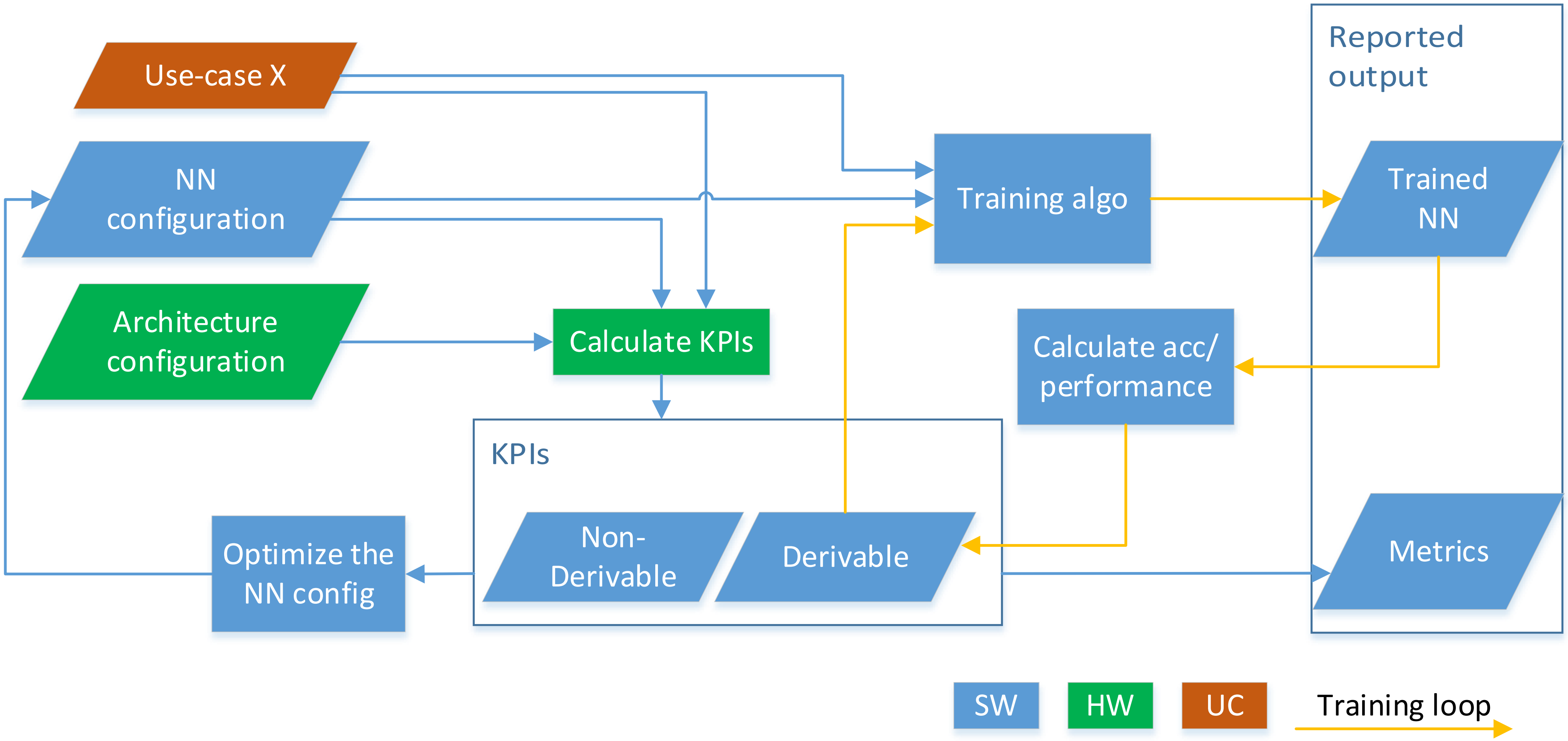
Example Accelerator



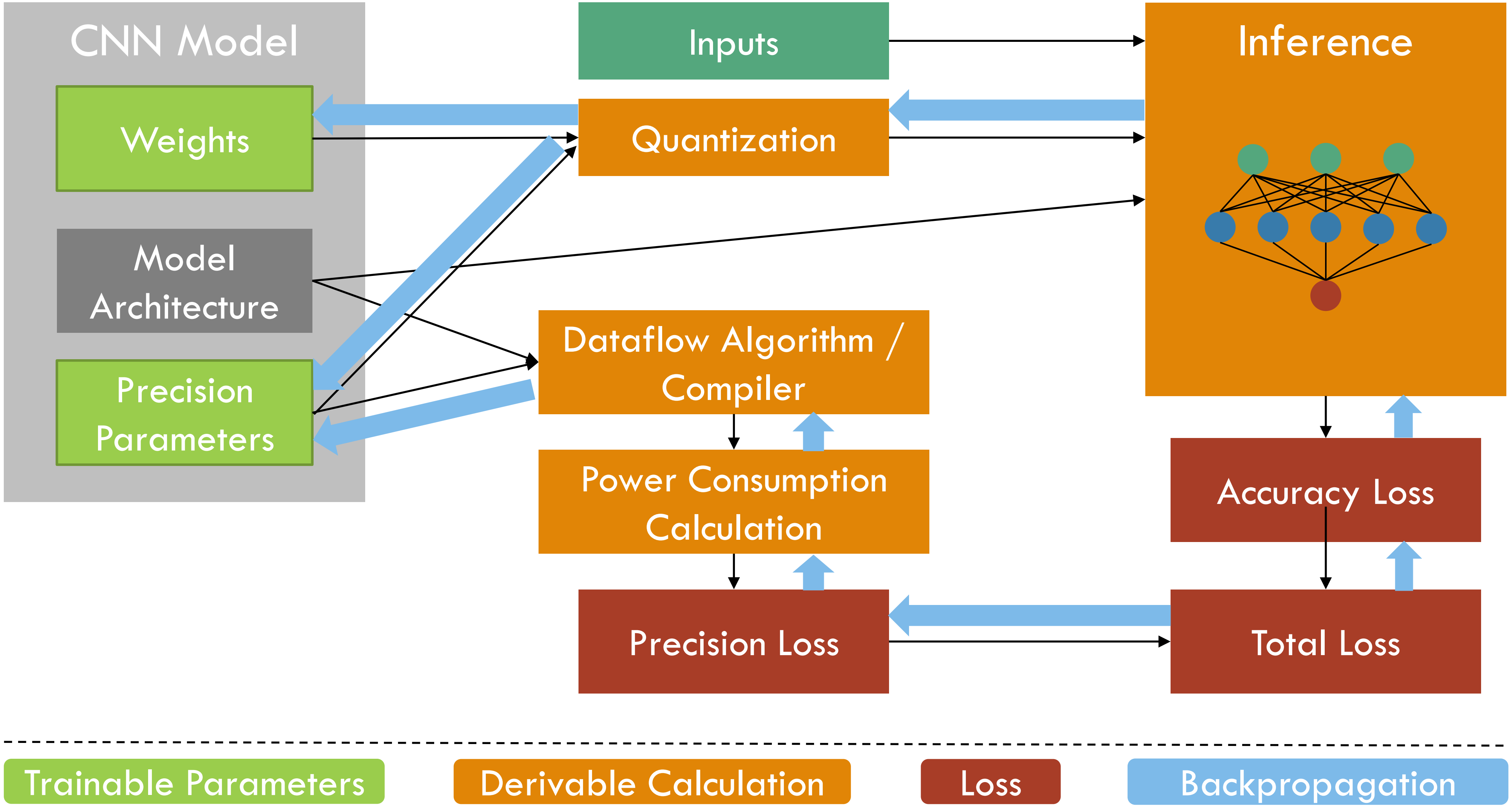
Abstract Architectures



PROPOSED TRAINING FLOW



EXAMPLE FOR KPI EVALUATION



CONSTRAINTS OF TRAINING FLOW

- Derivable KPIs integrated into training algorithm
- Derivable and Non-Derivable KPIs used for NN optimization
  - Manual: NN engineer designs better network
  - Automatic: Meta-learning using RL, evolutionary algorithms, random or grid search
- Main bottleneck is NN training time to measure accuracy
  - Reduced by retraining modified networks
- Trade off between metrics:
  - Use-case level specification
  - Hard thresholds and weighted average. E.g:

$$F(a,p,l,t) = \begin{cases} \infty, & \textbf{\textit{if}}\ \ a < 90\% \ \ \textbf{\textit{or}}\ \ p > 1mW \ \ \textbf{\textit{or}}\ \ l > 50ms \\ w_a a + w_p(1-p) + w_l(50-l) + w_t t, & \textbf{\textit{otherwise}} \end{cases}$$

ADVANTAGES AND APPLICATIONS

- Main benefits:
  - HW accelerators are benchmarked (normalized) on **use-case** basis not NN models
  - Different HW accelerators compared according to target applications
  - HW-aware training: Optimize trained NN for target HW
- Side benefits:
  - Design space exploration: HW optimized using different HW configurations
  - Non and -/derivable KPIs can enhance Meta-learning to find best NN for target HW (HW-aware Neural Architecture Search)
  - The developer knows (at early stage) metrics of proposed NNs

REQUIRED ACCELERATOR SIMULATIONS

Model	Description	Required functions	Optional functions
Action count calculation / NN Mapping	Action count calculation given ANN/SNN workload	• Action count calculation based on accelerator specific dataflow algorithm	• Dataflow optimization • Configurable dataflow algorithm
KPI evaluation	Evaluation of KPIs given action count	• KPI calculation for given accelerator (power, latency, hardware-aware accuracy etc.)	• Design space explorations with configurable accelerator model • Integration in backprop for derivable calculation
NN Constraints	Supported NN layers, parameter precision, etc.	• All supported hardware functions are provided	

Don't reinvent the wheel → Timeloop and Accelergy (Tutorial), NeuroSim, Netadapt

CONCLUSIONS

- New proposal for hardware optimization supporting benchmarking
  - ✓ Accelerators benchmarked for **use-cases** not NN models
  - ✓ HW-aware training: Optimize the trained NN for target HW
  - ✓ Optimization for specific KPIs and use cases
- Requirements, constraints and advantages
- This idea might contribute to standardization of fair benchmarking
- Partners must provide software models of their accelerators

THANK YOU!

QUESTIONS?

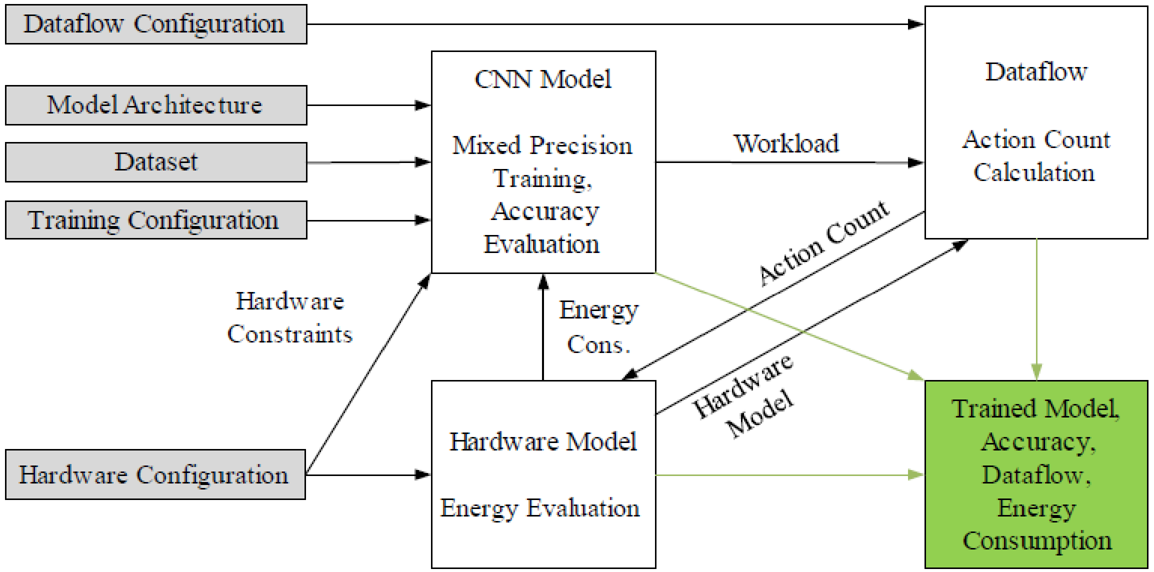
BACK-UP SLIDES

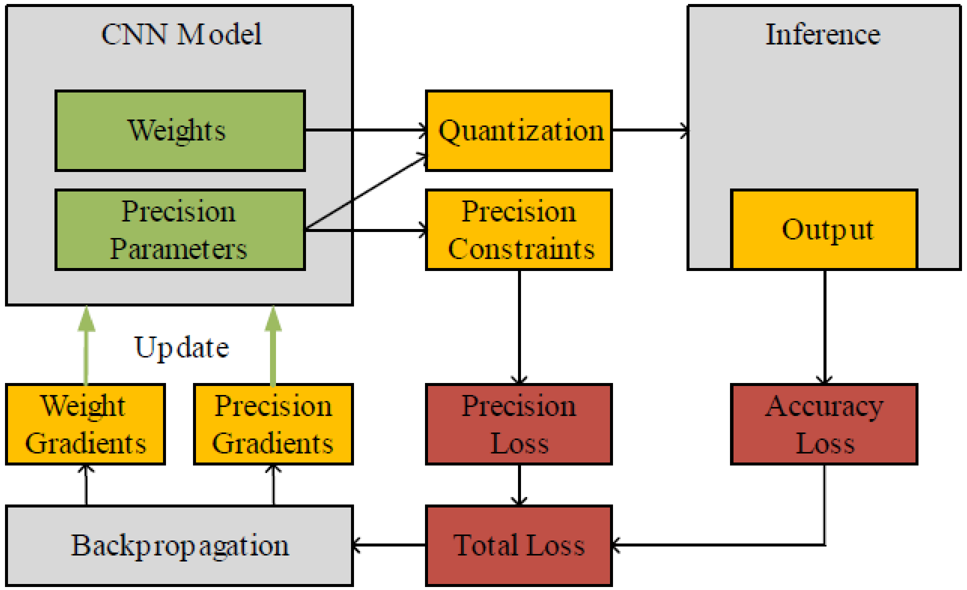
KEY PERFORMANCE INDICATORS

- Important KPIs to track (specific for each use-case)
  - Important: Accuracy, Power and Latency
  - Optional: Throughput, Robustness
  - HW specific: Cost, Flexibility, Scalability
- How to combine/prioritize:
  - Use-case level specification
  - Hard thresholds and weighted average:

$$F(a,p,l,t) = \begin{cases} \infty, & if\ p > 1mW\ or\ l > 50ms \\ w_a a + w_p(1-p) + w_l(50-l) + w_t t, & otherwise \end{cases}$$

TRAINING FOR DERIVABLE KPI CALCULATIONS

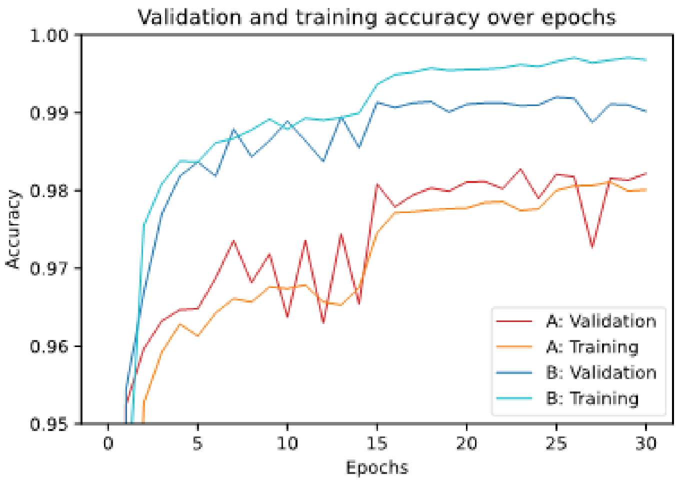


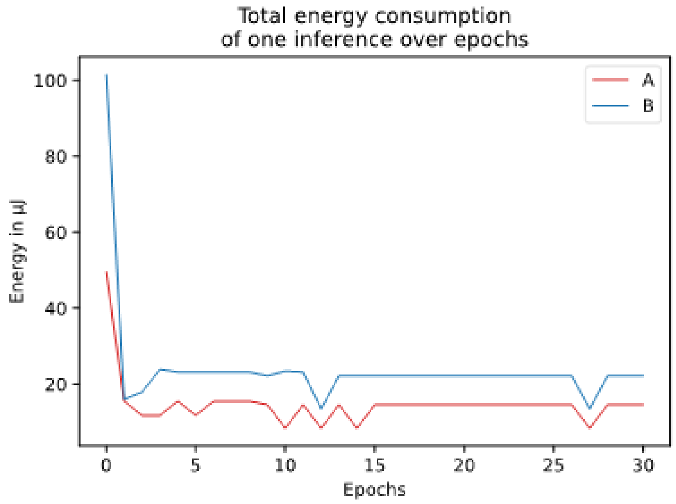


EXAMPLE KPI-AWARE TRAINING

- Algorithm implemented in TensorFlow
- Trained on MNIST with two small CNN models
- Full precision accuracy with SGD (B): 99.31%
- Mixed precision accuracy with SGD (B): 99.20%

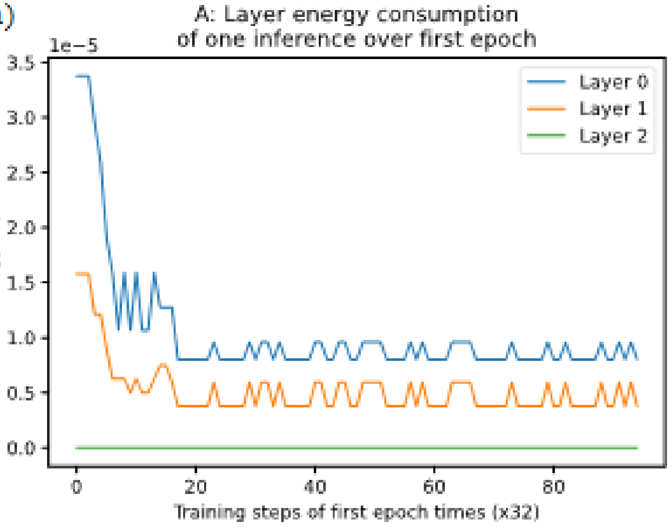
Model Configurations	
A	B
3 weight layers	6 weight layers
input (28 × 28 grayscale image)	
conv3-16	conv3-16 conv3-16
maxpool	
conv3-32	conv3-32 conv3-32
maxpool	
dropout FC-10	FC-128 dropout FC-10



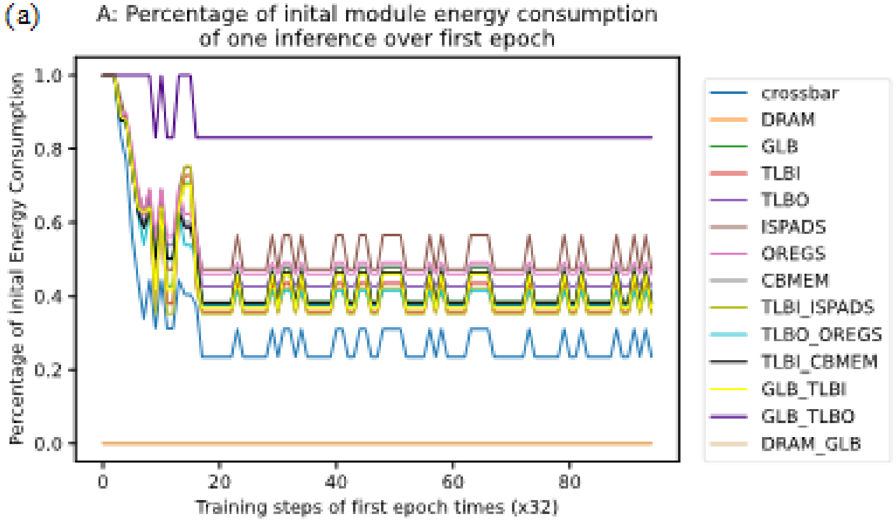


LAYER AND MODULE ENERGY

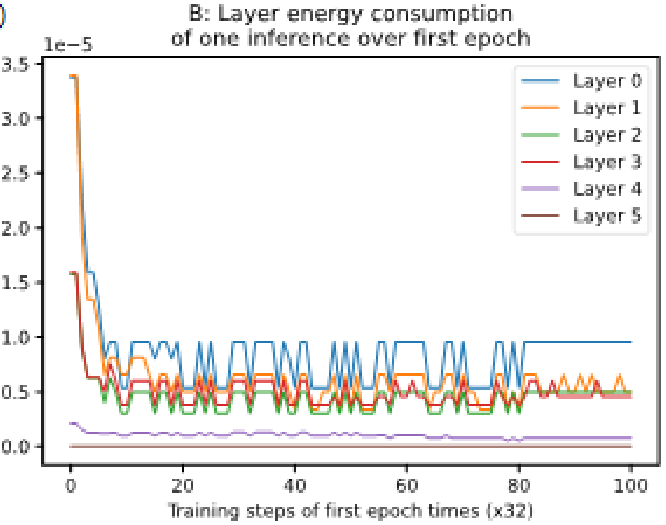
(a)



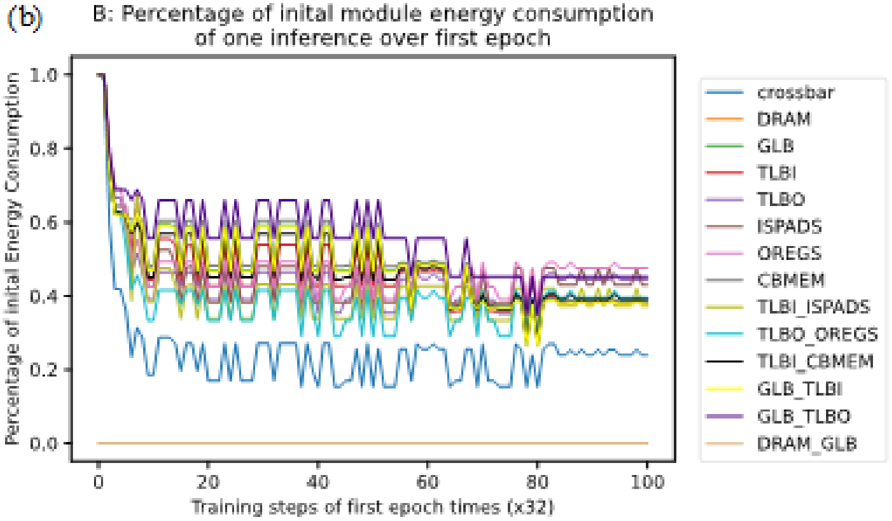
(a)



(c)



(b)



LEARNED PRECISIONS