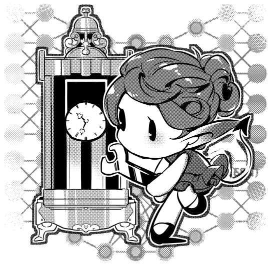
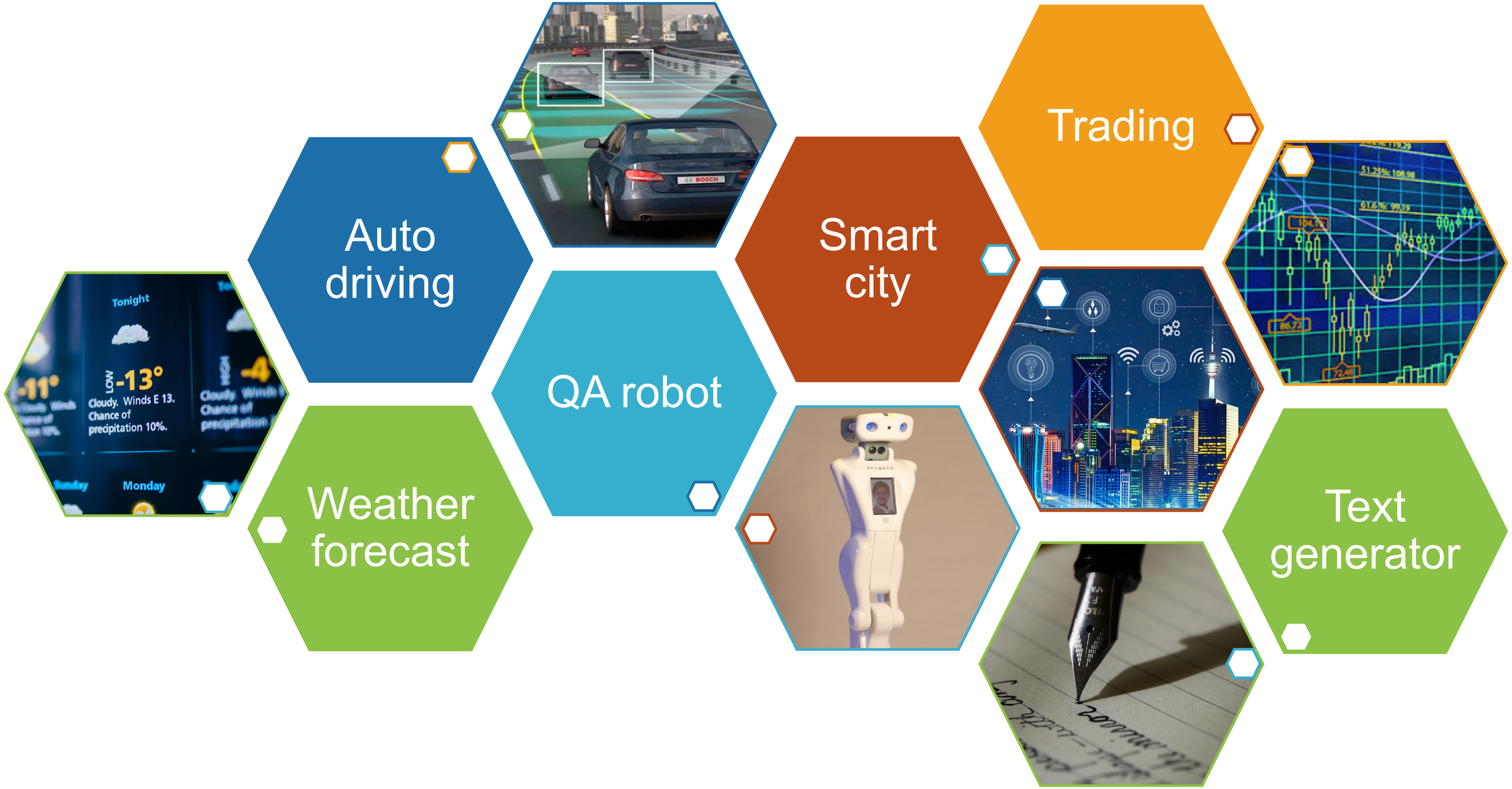
ALERT: Accurate Learning for Energy and Timeliness

Chengcheng Wan, Muhammad Husni Santriaji,
Eri Rogers, Henry Hoffmann, Michael Maire and Shan Lu

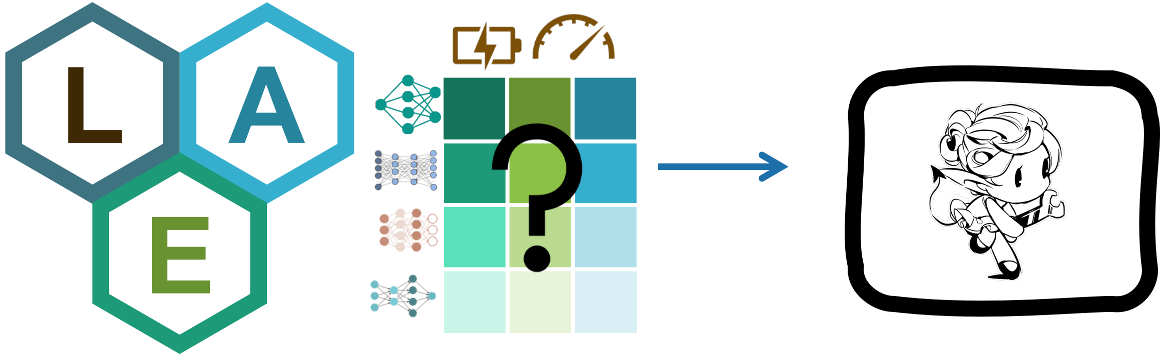
UCHICAGO

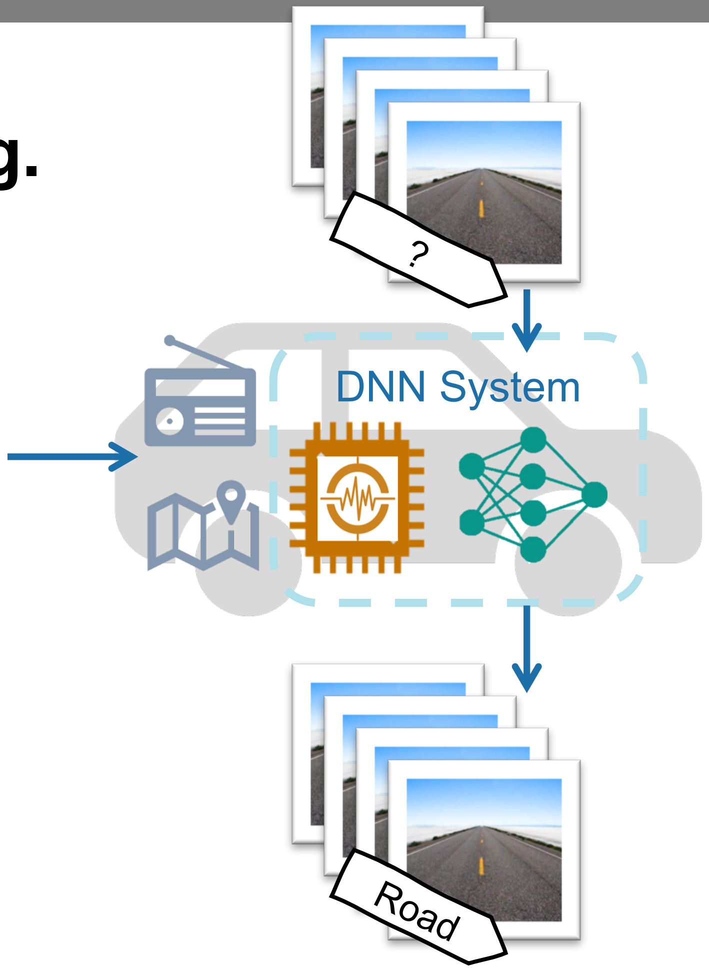


DNN is Deployed Everywhere



DNN Deployment is Challenging.

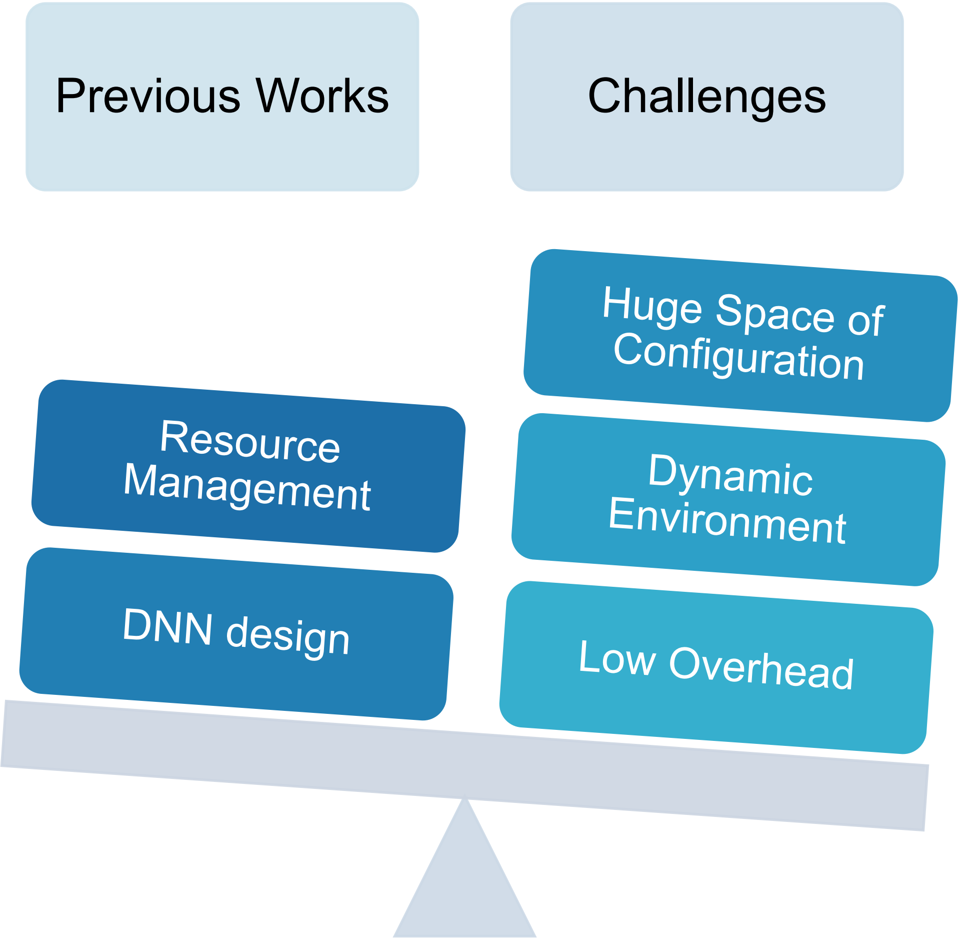




Challenges

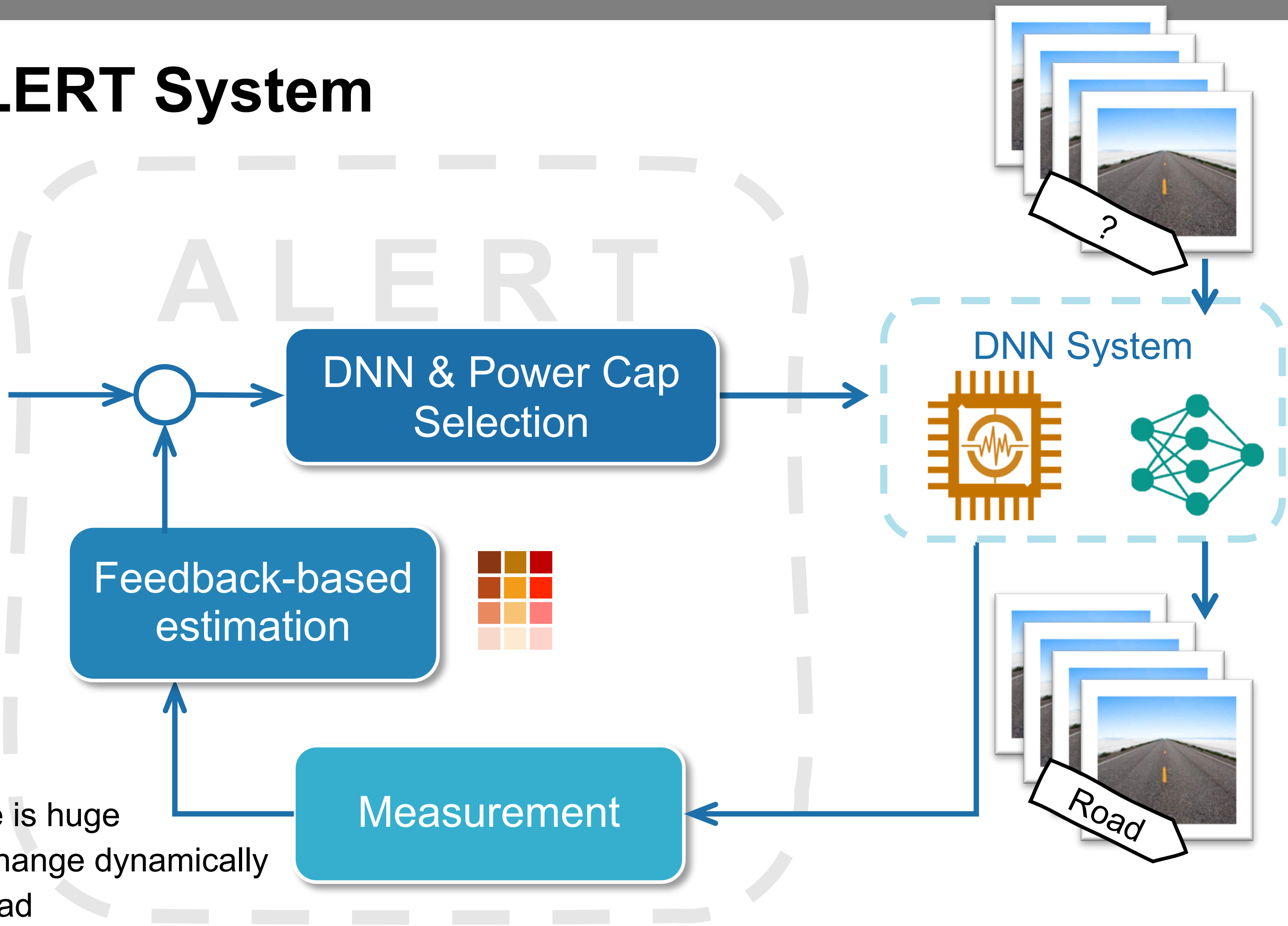
- Configuration space is huge
- Environment may change dynamically
- Must be low overhead

Previous Work



[1] H. Hoffmann et. al. Jouleguard:
energy guarantees for approximate applications. SOSP, 2015.
[2] C. Imes et. al. Poet: a portable approach to minimizing energy under soft real-time constraints. RTAS, 2015
[3] N. Mishra et. al. CALOREE: learning control for predictable latency and low energy. ASPLOS, 2018.
[4] A. Rahmani et. al. SPECTR: formal supervisory control and coordination for many-core systems resource management. ASPLOS, 2018.
…

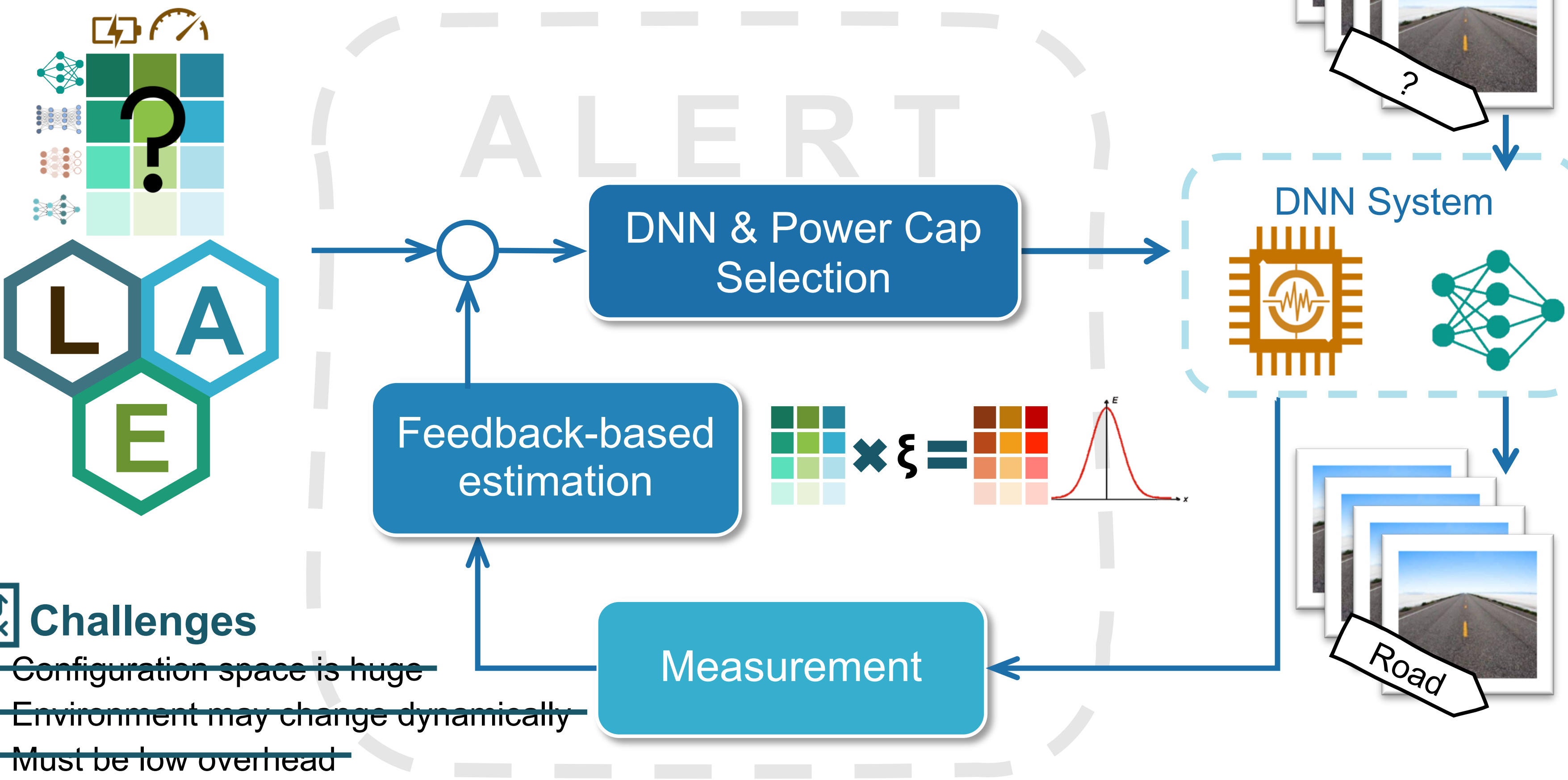
Our ALERT System



Challenges

- Configuration space is huge
- Environment may change dynamically
- Must be low overhead

Our ALERT System



Challenges

- ~~Configuration space is huge~~
- ~~Environment may change dynamically~~
- ~~Must be low overhead~~

Evaluation Highlights

✓ ALERT satisfies LAE constraints.

99.9% cases for vision; **98.5%** cases for NLP

✓ Probabilistic design overcomes dynamic variability efficiently.

ALERT achieves **93-99%** of Oracle's performance

✓ Coordinating App- and Sys- level improves performance.

Reduces **13%** energy and **27%** error over prior approach

Outline

- Understanding DNN Deployment Challenges
- ALERT Run-time Inference Management
- Experiments and Results

Outline

- Understanding DNN Deployment Challenges
- ALERT Run-time Inference Management
- Experiments and Results

Experiment Settings

4 DNNs

ResNet50, VGG16, RNN, Bert

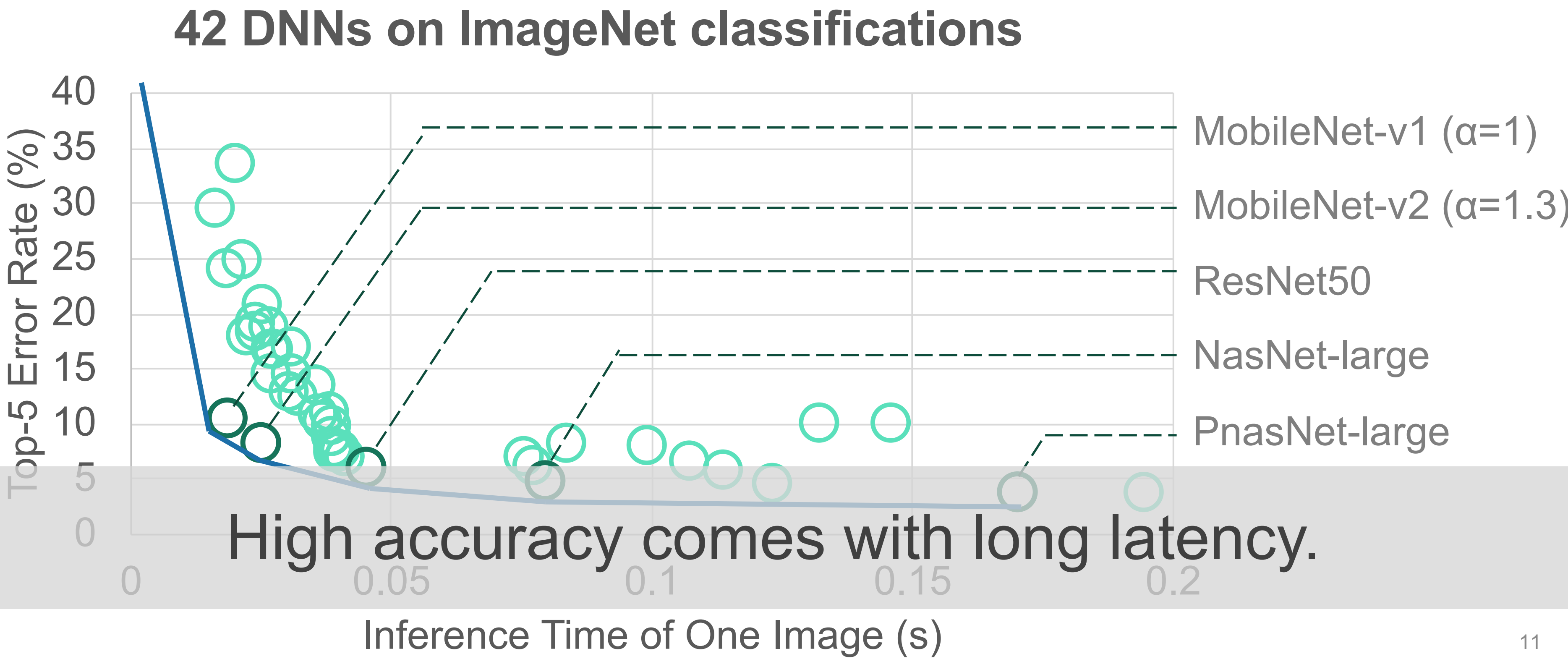
4 Platforms

ODroid, CPUs, GPU

3 Tasks

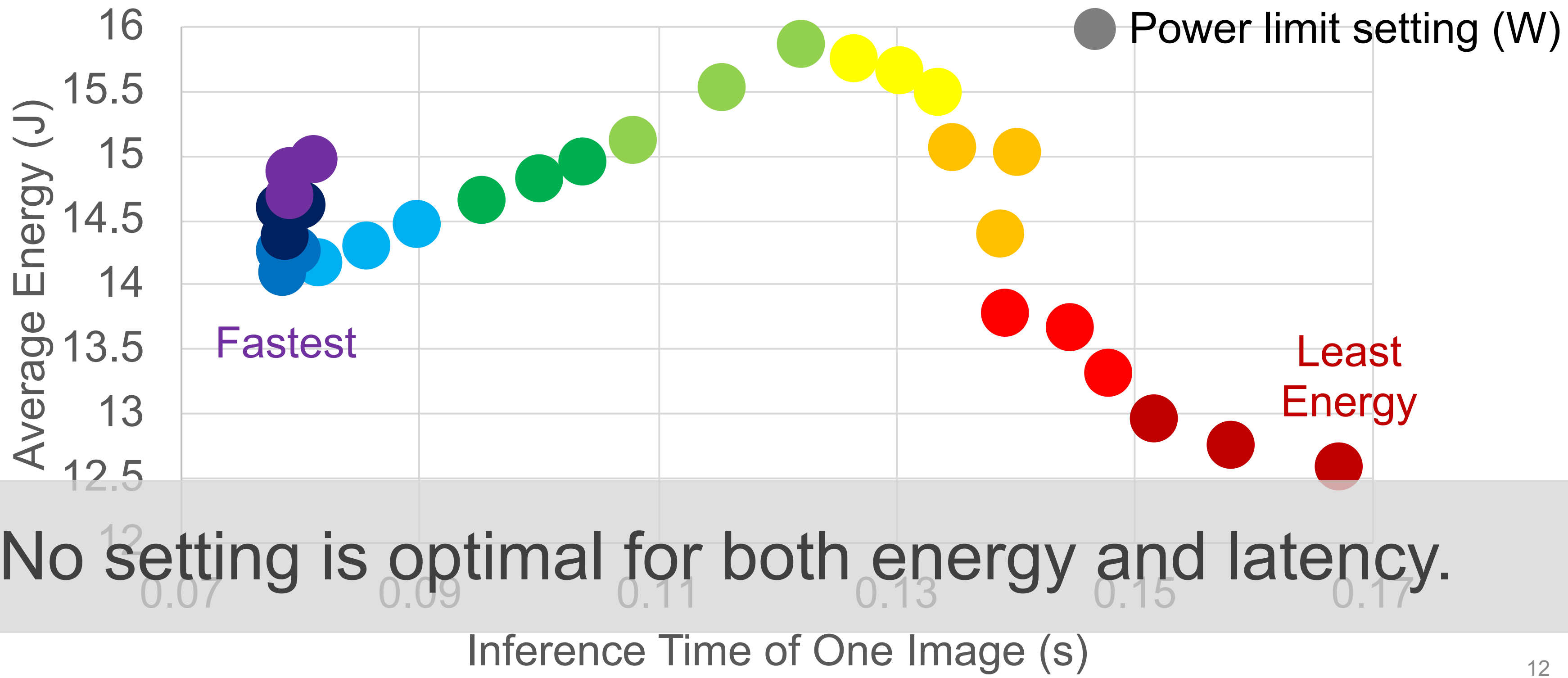
Image classification (ImageNet)
Sentence prediction (PTB)
Question Answering (SQuAD)

Tradeoffs from DNNs



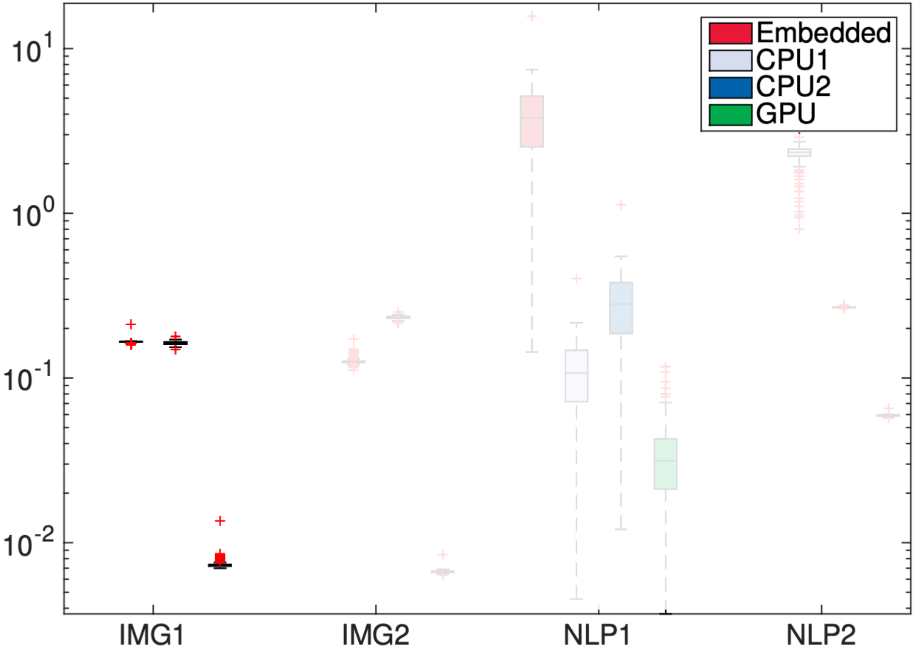
High accuracy comes with long latency.

Tradeoffs from System Settings

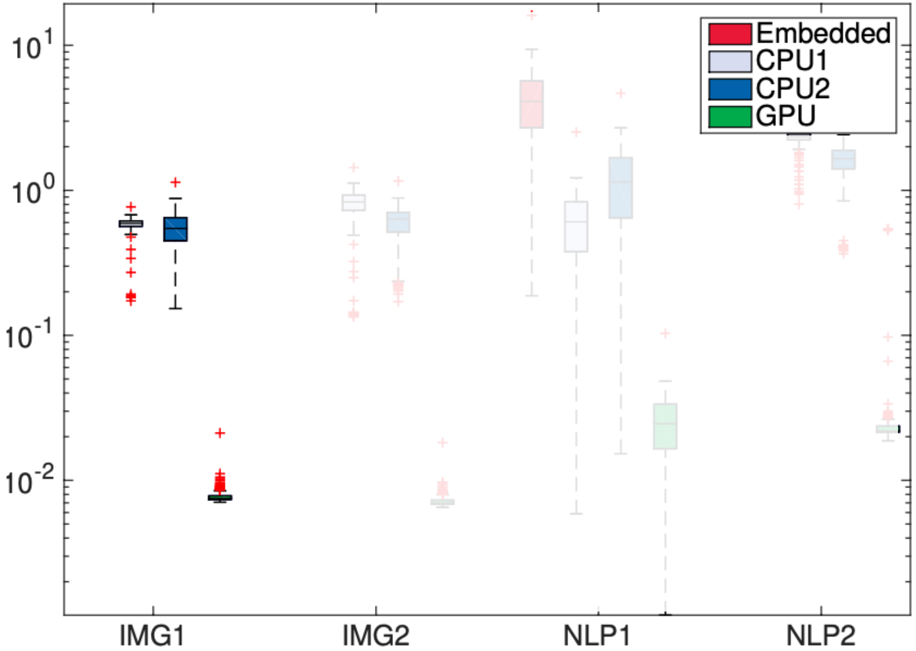


No setting is optimal for both energy and latency.

Run-time Variability

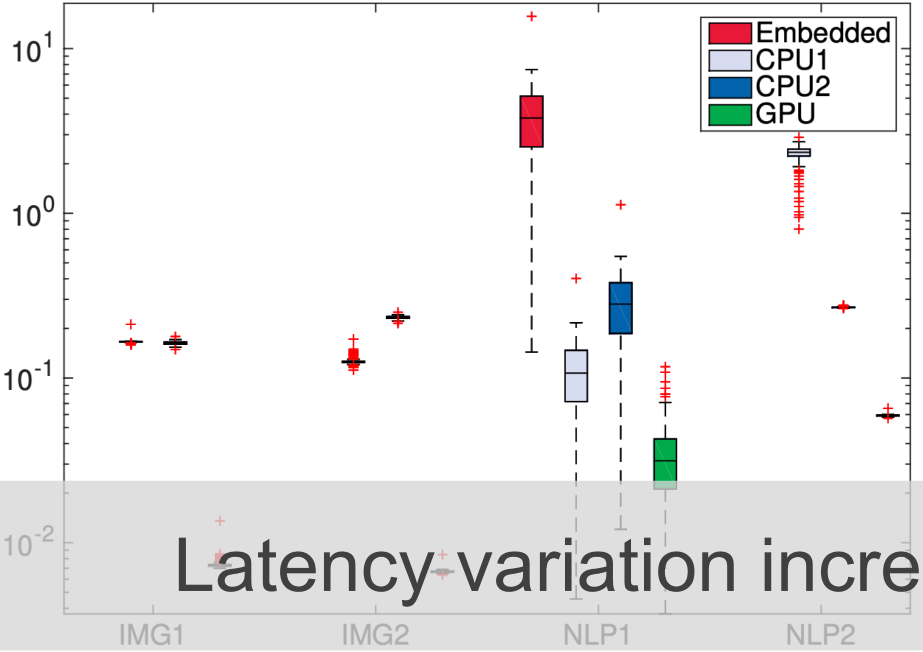


Without co-locate job

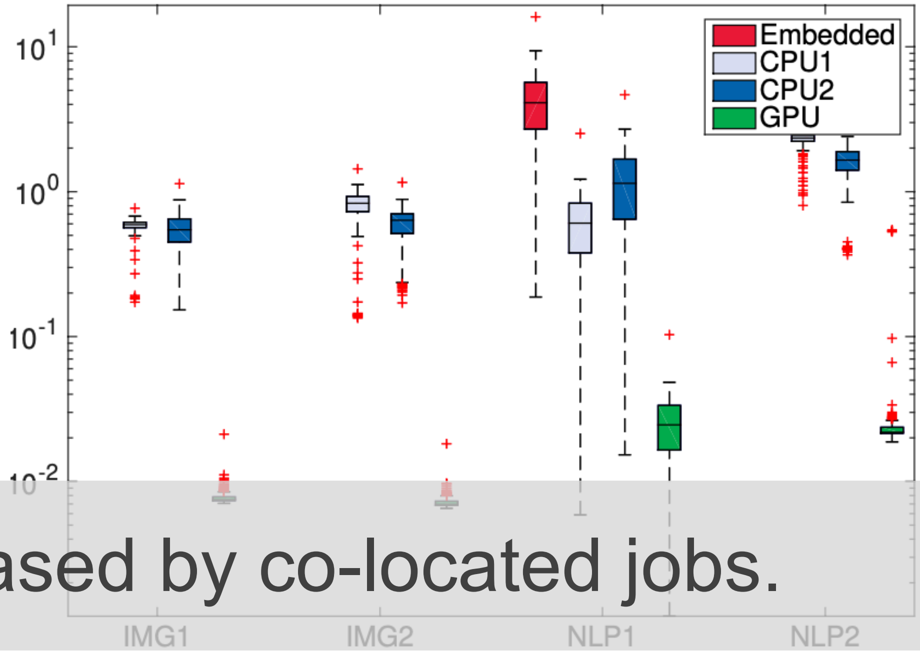


With co-locate job

Run-time Variability



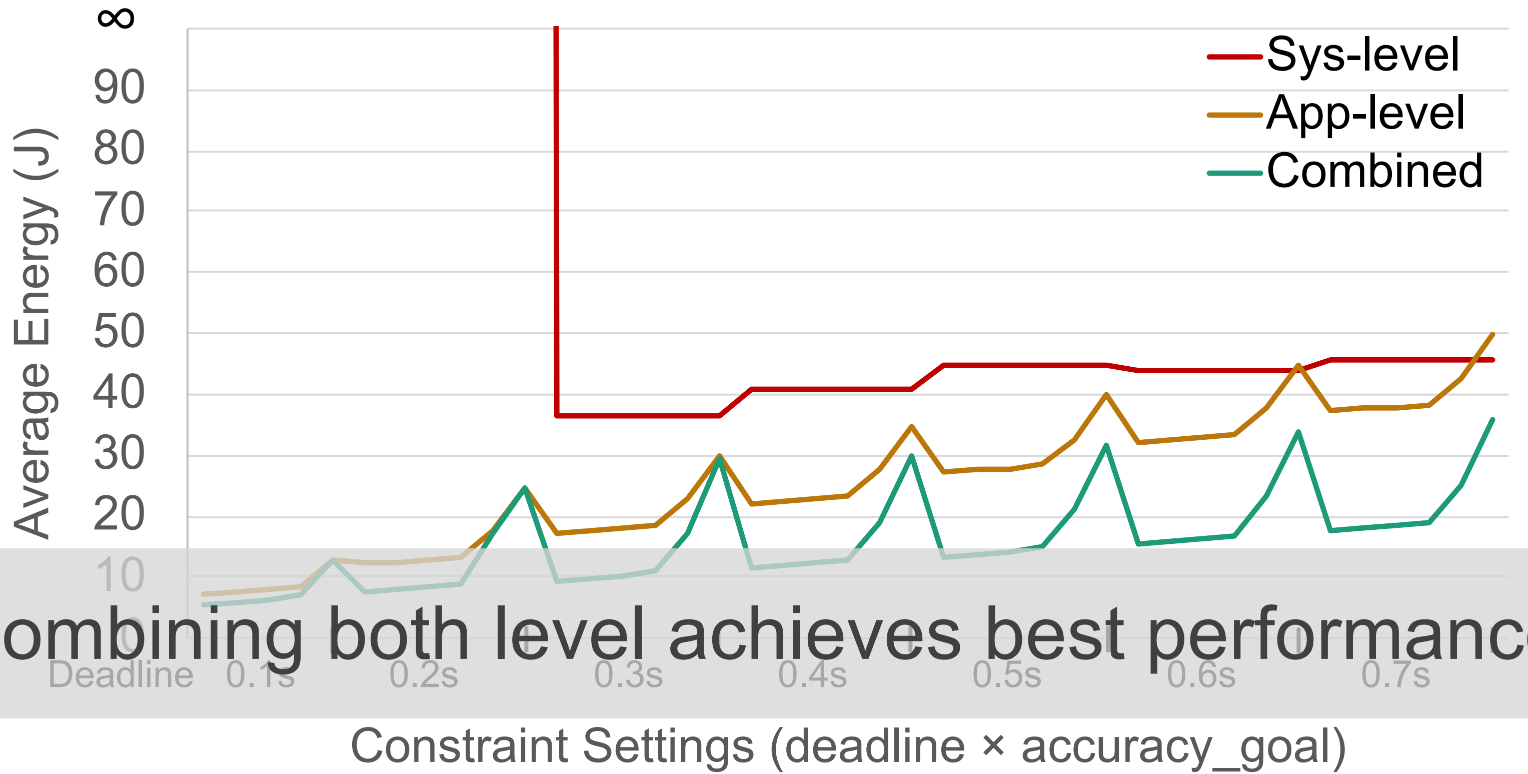
Without co-locate job



With co-locate job

Latency variation increased by co-located jobs.

Potential Solutions



Combining both level achieves best performance.

Outline

- Understanding DNN Deployment Challenges
- ALERT Run-time Inference Management
- Experiments and Results

Three Dimensions & Two Tasks

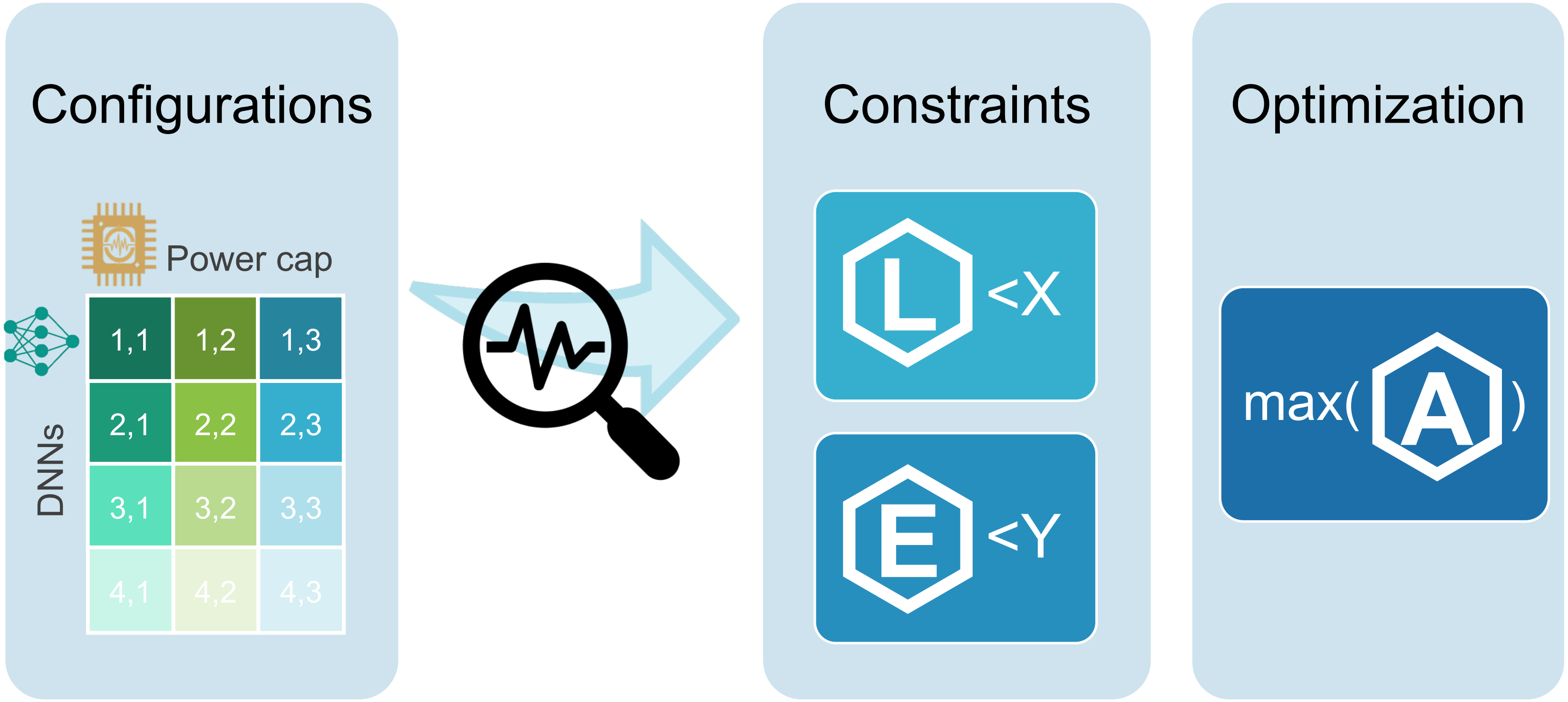
Maximize Accuracy

With energy consumption goal and inference deadline

Minimize Energy

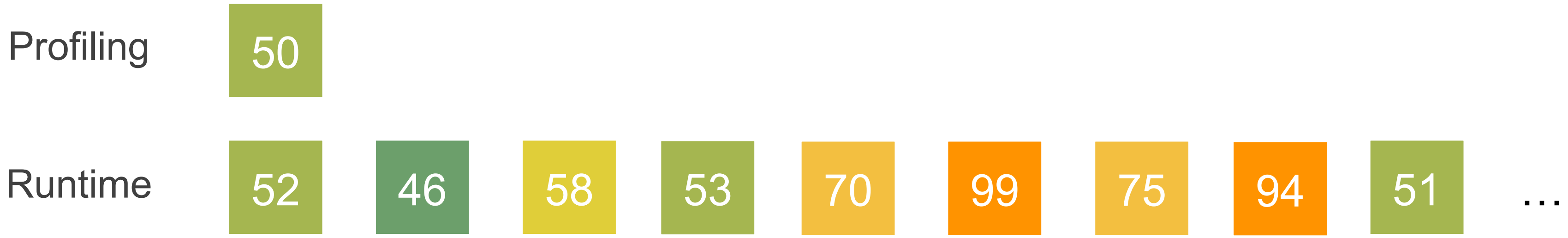
With accuracy goal and inference deadline

Maximize Accuracy Task



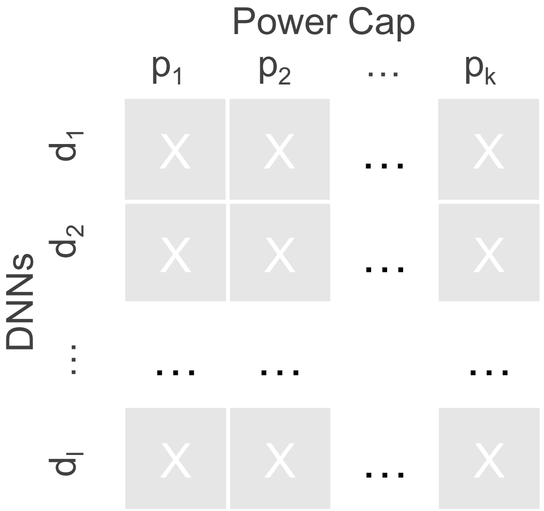
How to estimate the inference latency?

- Two key challenges
  - Runtime variation: The inference time may be different even for same the configuration



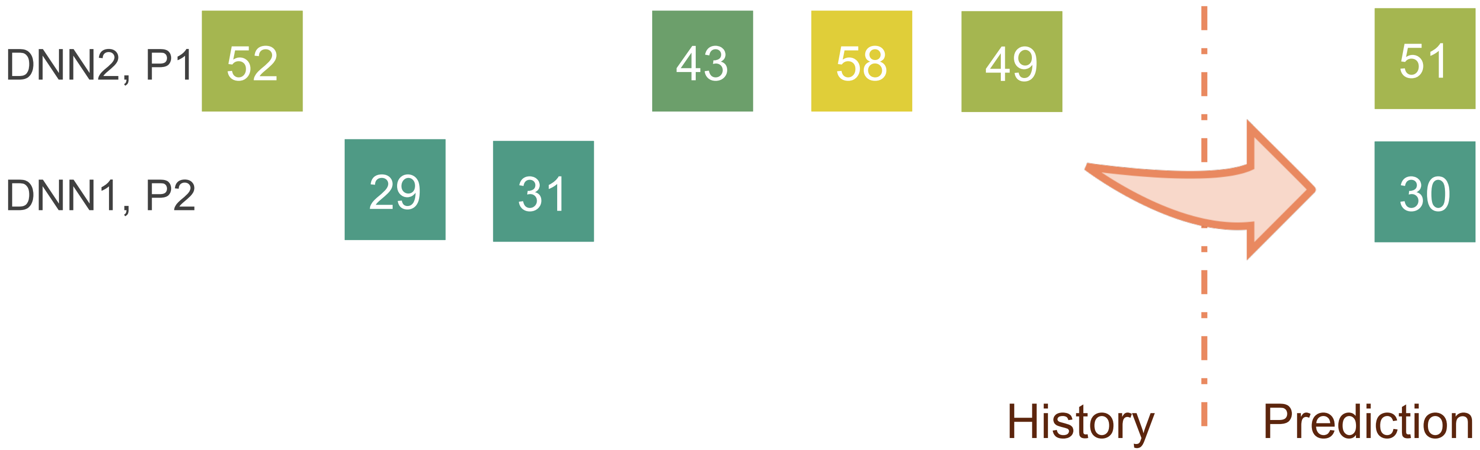
How to estimate the inference latency?

- Two key challenges
  - Runtime variation
  - Too many combinations of DNNs and resources



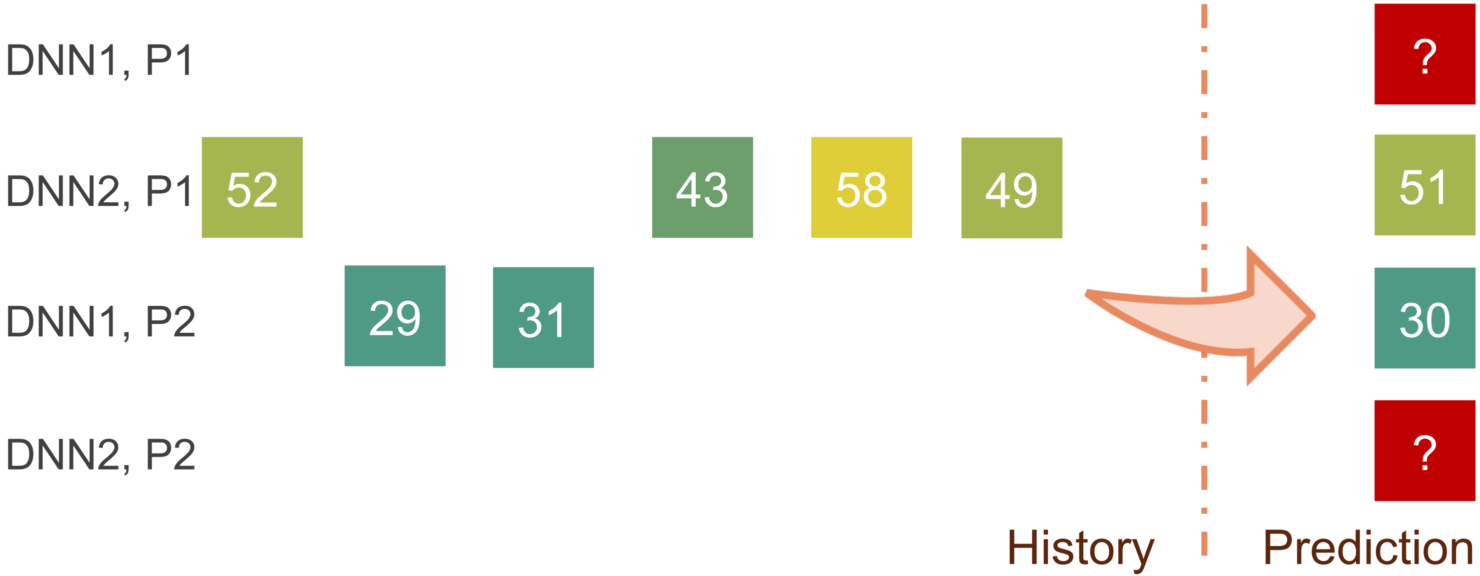
Potential Solution

- Kalman filter
  - Estimate latency for each configuration
  - Use recent execution history



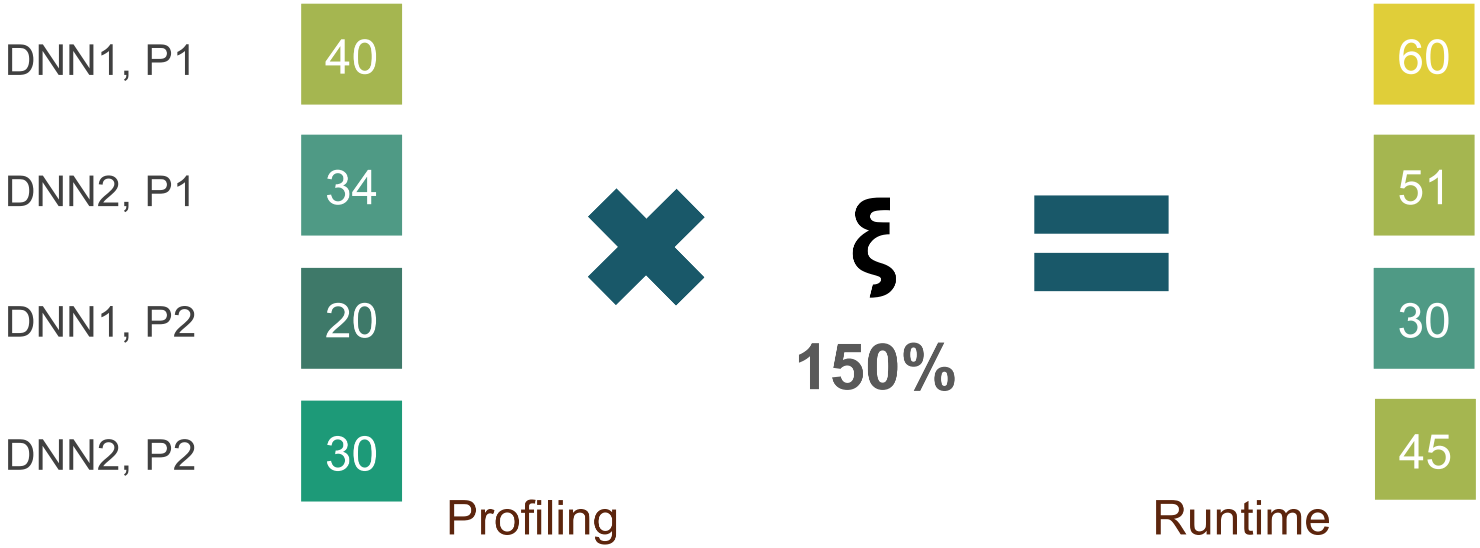
Potential Solution: drawback

- Cannot solve the problem
  - Not enough history for each configuration



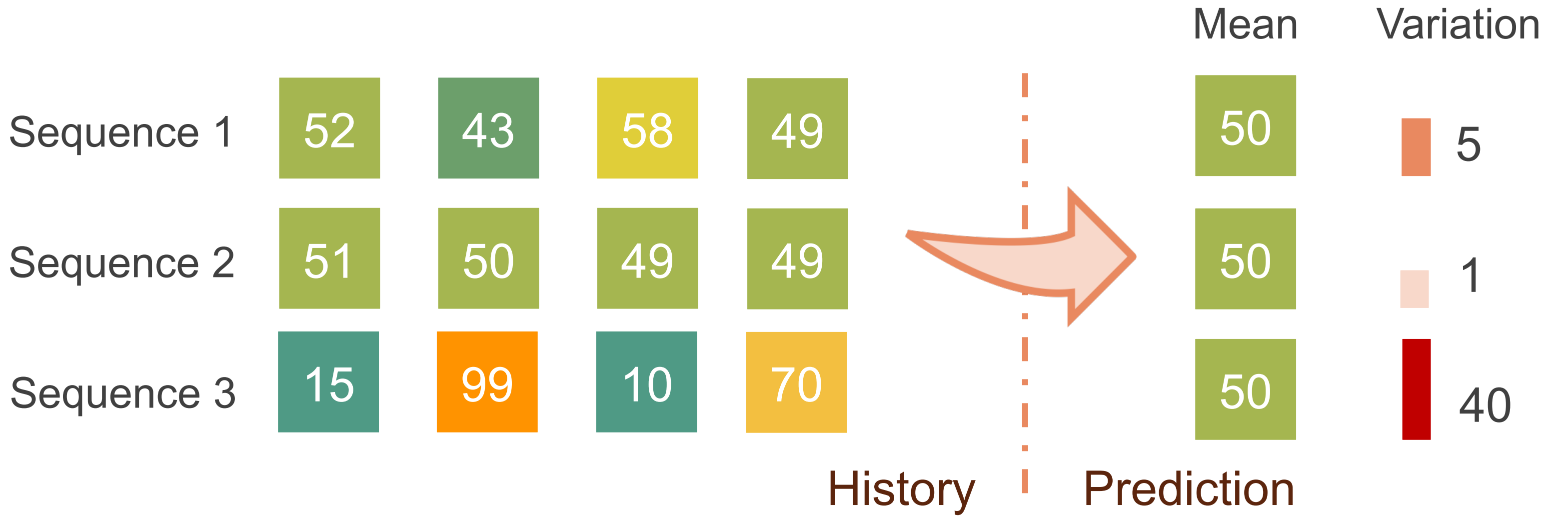
How to estimate the inference latency?

- Global Slow-down factor ξ
  - Use recent execution history under **any** DNN or resources



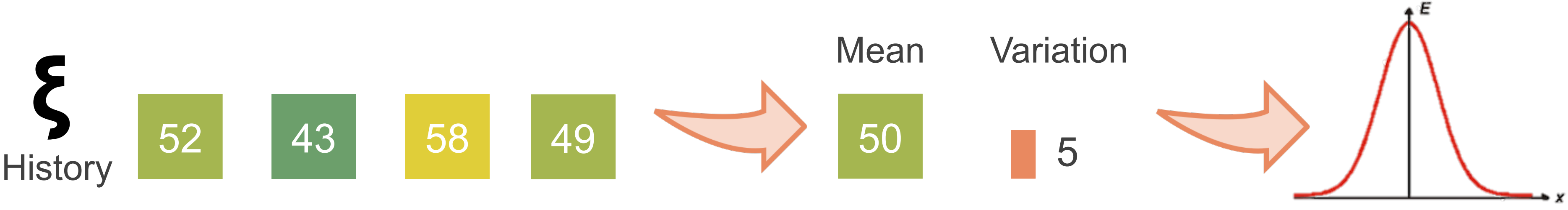
How to estimate the inference latency?

- Mean estimation is not sufficient
  - The variation might be too big to provide a good prediction.
- Different implications on DNN selection



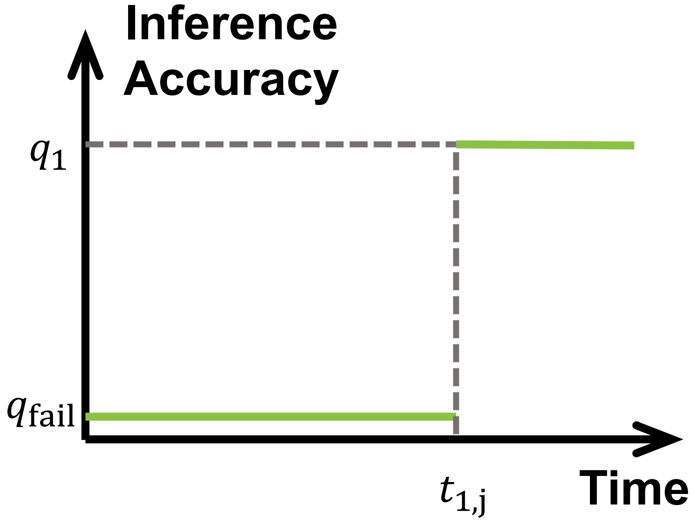
How to estimate the inference latency?

- Global Slow-down factor ξ
  - Use recent execution history under **any** DNN or resources
  - Estimate its distribution: mean and variance

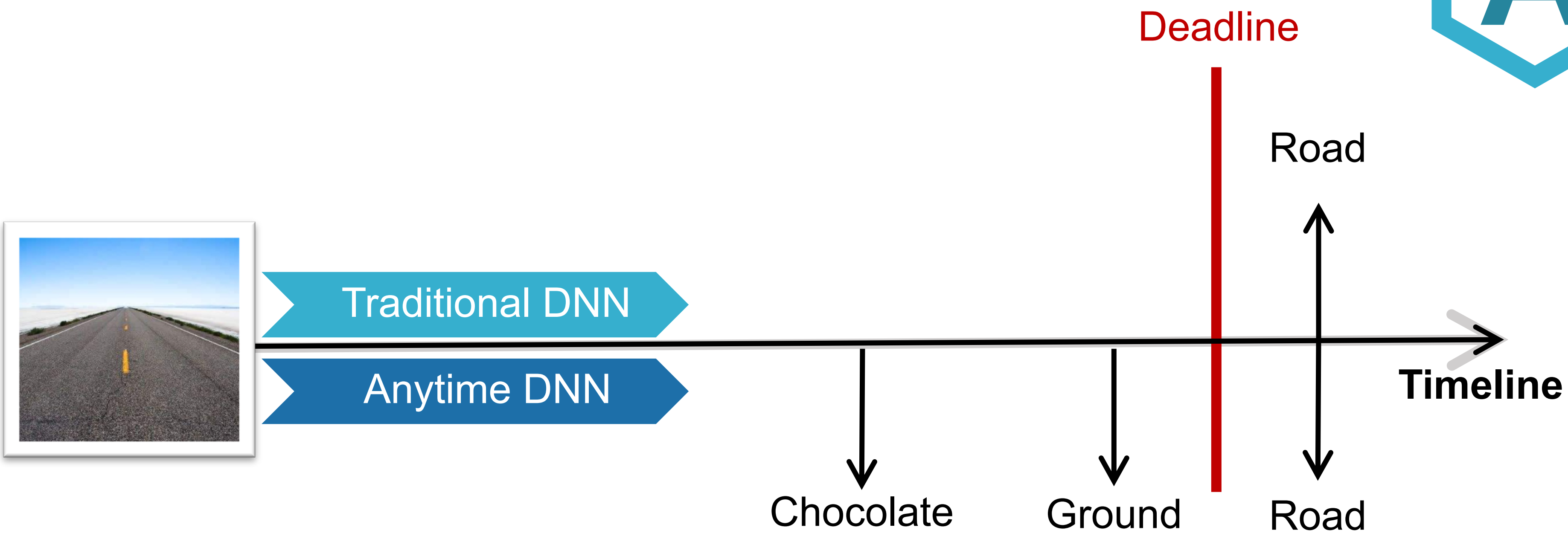


How to estimate accuracy under a deadline?

- Can inference be finished before deadline?
  - If yes, training accuracy of the selected DNN
  - If not, random guess accuracy
    - Unless it’s an Anytime DNN.



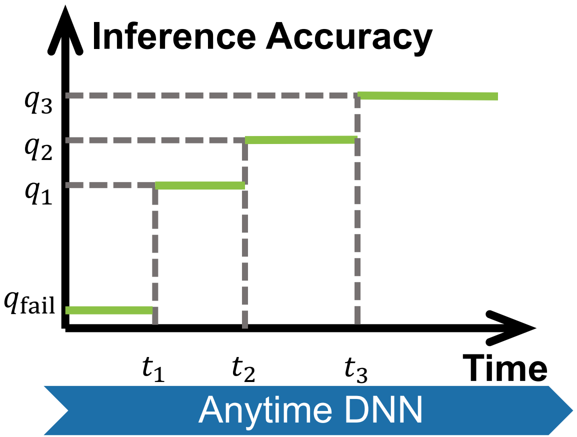
What is an Anytime DNN?

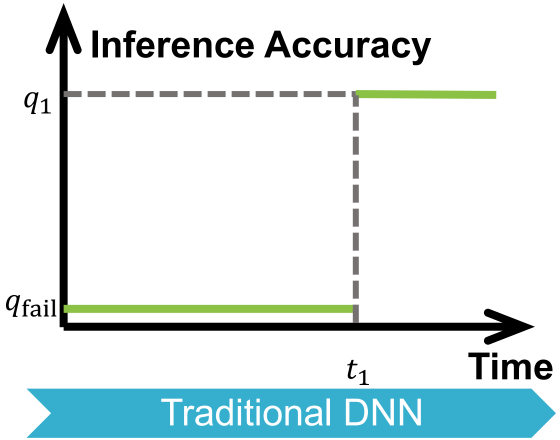


[1] C. Wan et. al. Orthogonalized SGD and Nested Architectures for Anytime Neural Networks . ICML, 2020.

How to estimate accuracy under a deadline?

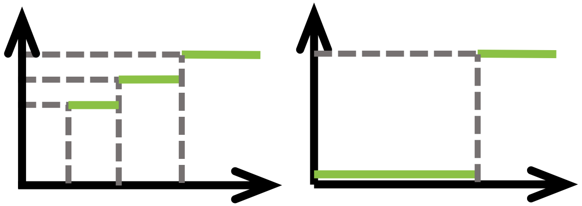
- Can inference be finished before deadline?
  - If yes, training accuracy of the selected DNN
  - If not,
    - Traditional DNN: random guess accuracy.
    - Anytime DNN: accuracy of the last output

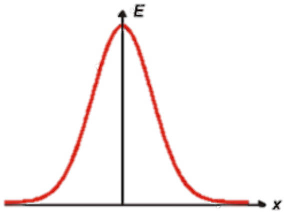




How to estimate accuracy under a deadline?

Accuracy-Latency Relation + Latency Distribution = Expectation of Accuracy

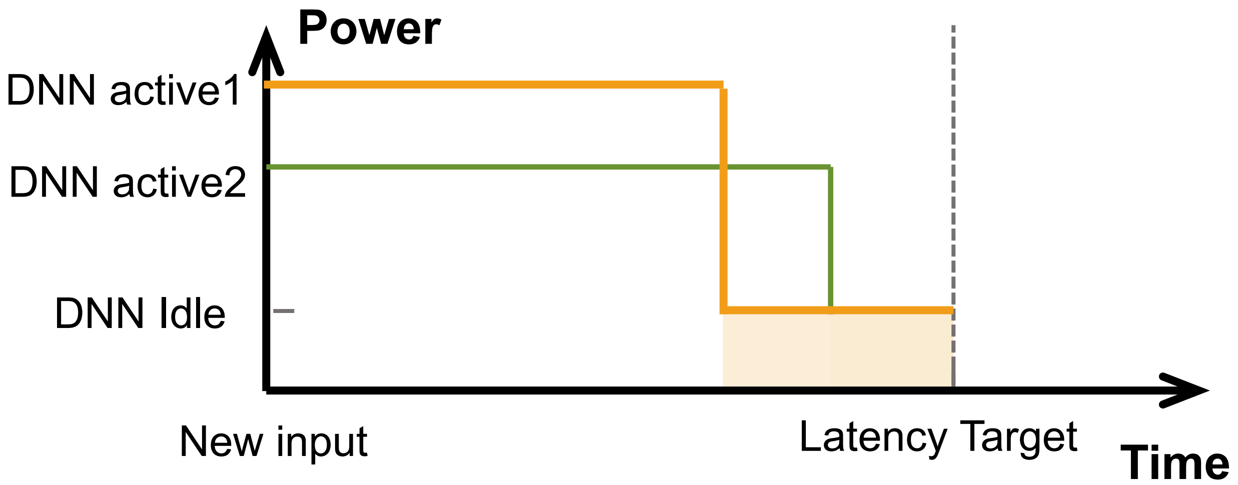




E(A)

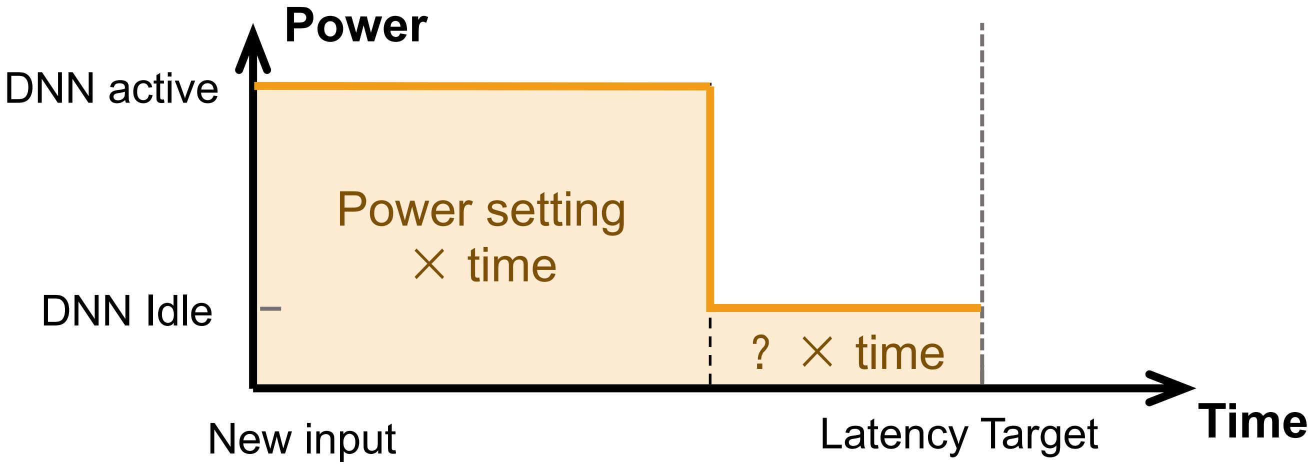
How to manage energy?

- Power-cap as a knob to configure system resource
- Idle power: other process may still consume energy when DNN inference has finished

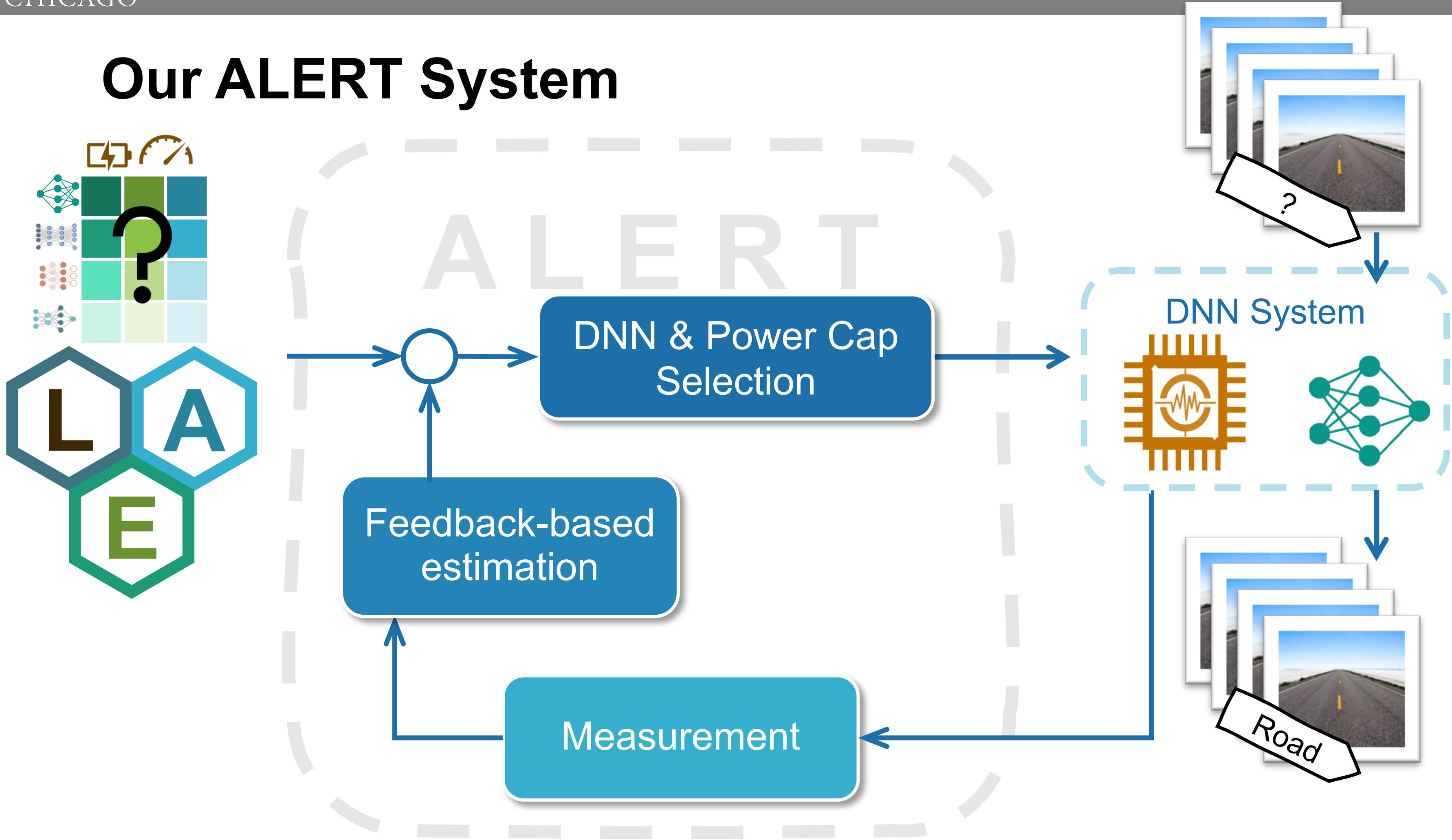


How to estimate the energy consumption?

- Estimate energy from power
  - DNN active power is power setting
  - DNN idle power is estimated by Kalman filter



Our ALERT System



Outline

- Understanding DNN Deployment Challenges
- ALERT Run-time Inference Management
- Experiments and Results

Experiment Settings

3 Platforms

CPUs, GPU

2 Tasks

1. Minimize energy
2. Maximize accuracy

2 DNNs

Sparse ResNet50, RNN

5 Scenarios

Default,
Compute intensive (2),
Memory intensive (2)

Schemes

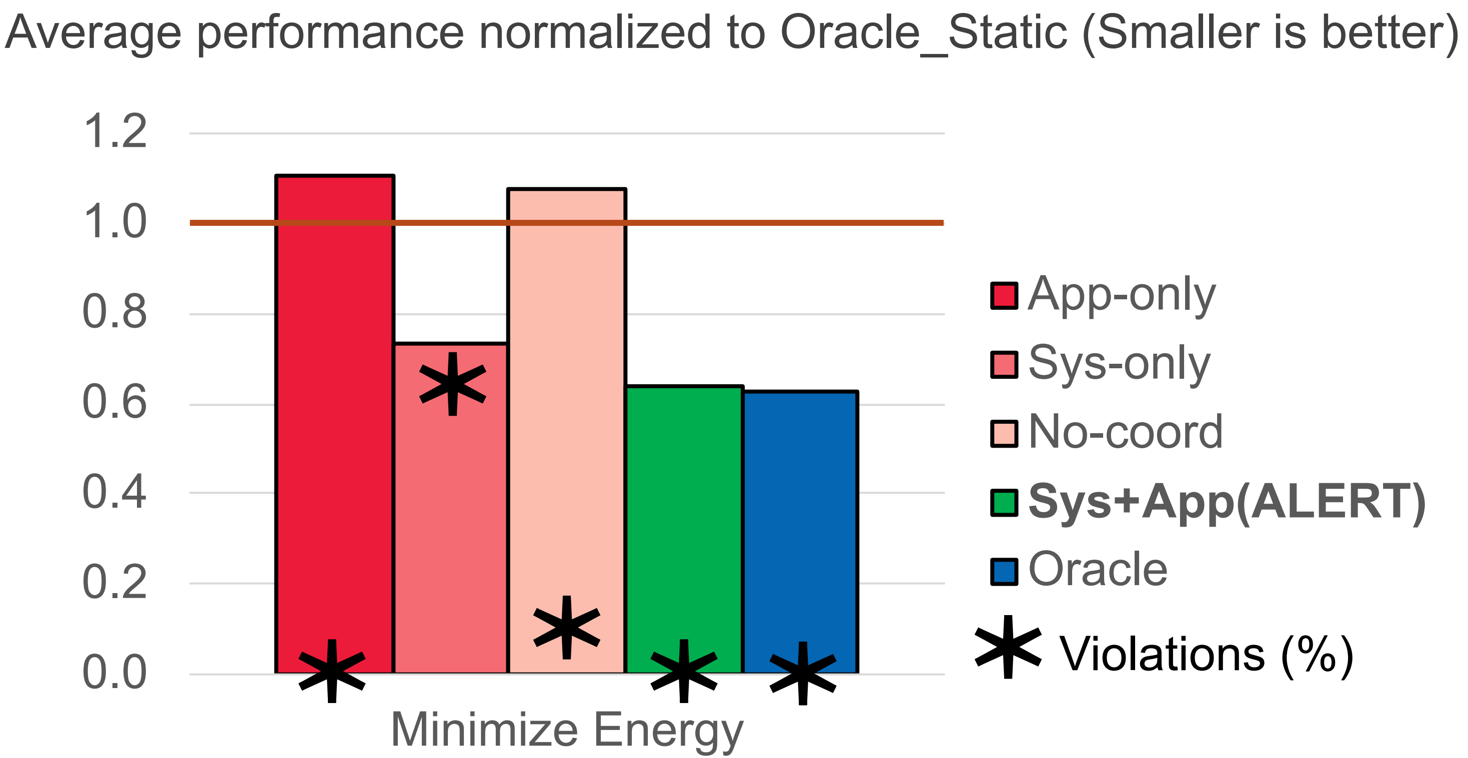
Oracles

- **Oracle**: Change configuration for every input. Assume perfect knowledge of future. Emulated from profiling result.
- **Oracle-static**: Same configuration for all inputs.

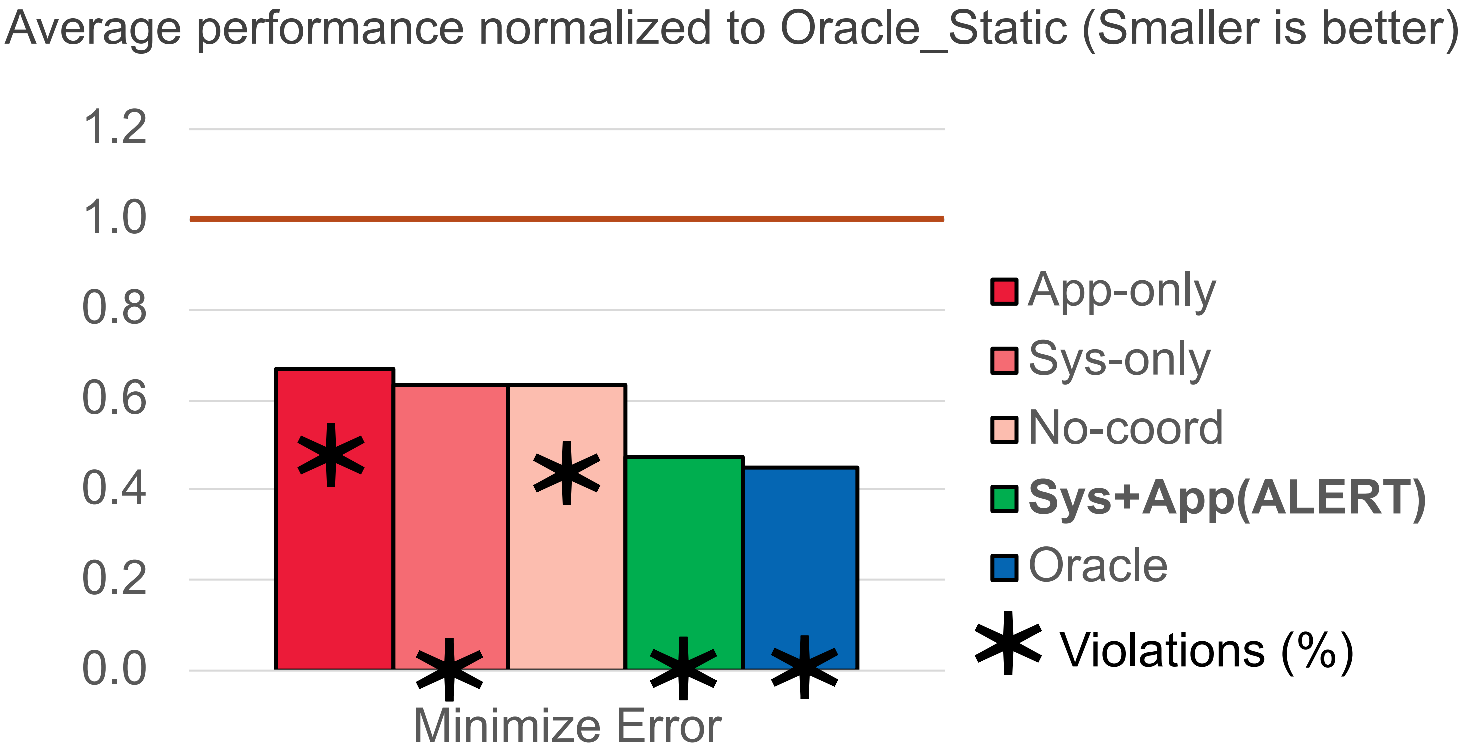
Baselines

- **Sys-only**: Only adjust power-cap
- **App-only**: Use an Anytime DNN
- **No-coord**: Anytime DNN without coordination with power-cap

Evaluation: Scheduler Performance

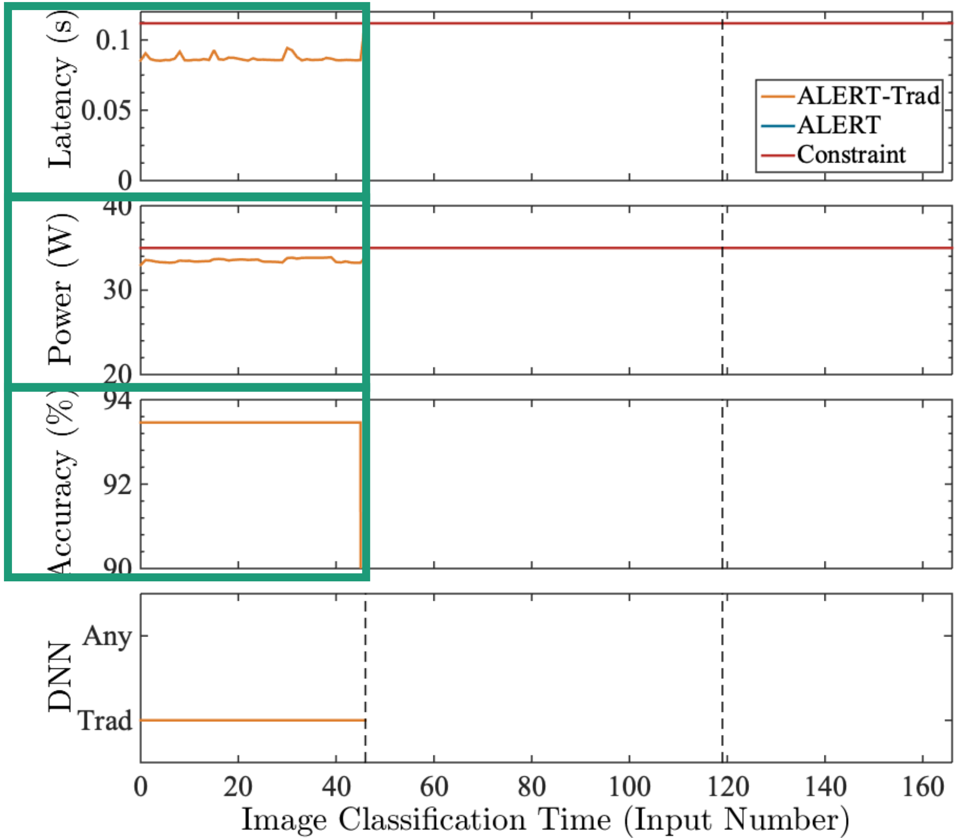


Evaluation: Scheduler Performance



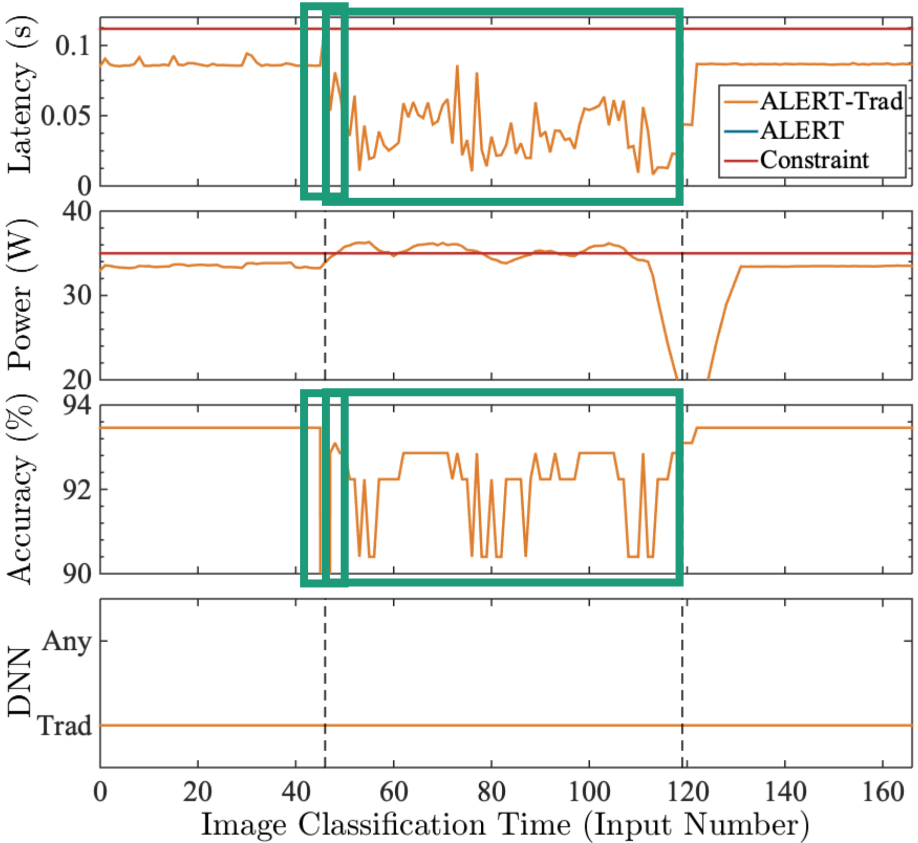
How ALERT Works with Traditional DNN

- Meet requirements in most cases
- Quickly detect contention changes
- Use anytime DNN under unstable environment



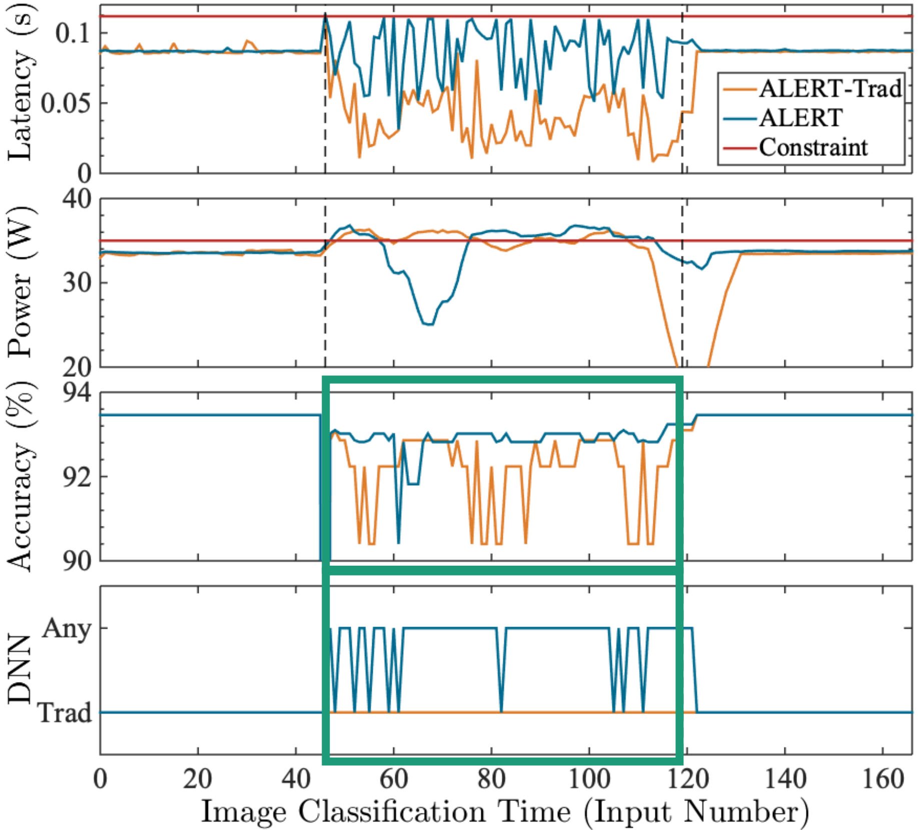
How ALERT Works with Traditional DNN

- Meet requirements in most cases
- Quickly detect contention changes
- Use anytime DNN under unstable environment



How ALERT Works with Anytime +Traditional DNN

- Meet requirements in most cases
- Quickly detect contention changes
- Use anytime DNN under unstable environment



Conclusion

- Understand DNN inference challenges
- ALERT Run-time inference System
- High performance and energy efficiency