DESIGN, AUTOMATION & TEST IN EUROPE
9 – 13 March, 2015 · Grenoble · France
The European Event for Electronic System Design & Test

Workload Uncertainty Characterization and Adaptive Frequency Scaling for Energy Minimization of Embedded Systems

Dr. Anup Das (University of Southampton)

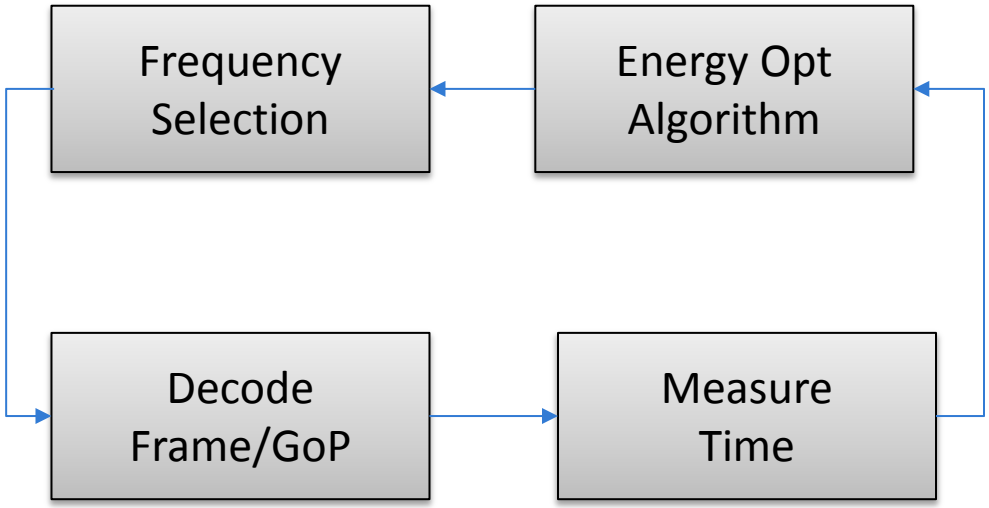
Co-authors:
Dr. Akash Kumar, A/P Bharadwaj Veeravalli (National University of Singapore)
Dr. Rishad A. Shafik, A/P Geoff V. Merrett, Prof. Basir M. Al-Hashimi (University of Southampton)

Content

- Run-time Energy Optimization
- Workload Uncertainty
- Proposed Approach
  - Multinomial logistic regression
  - Maximum likelihood estimation
- Results

Run-time Energy Optimization

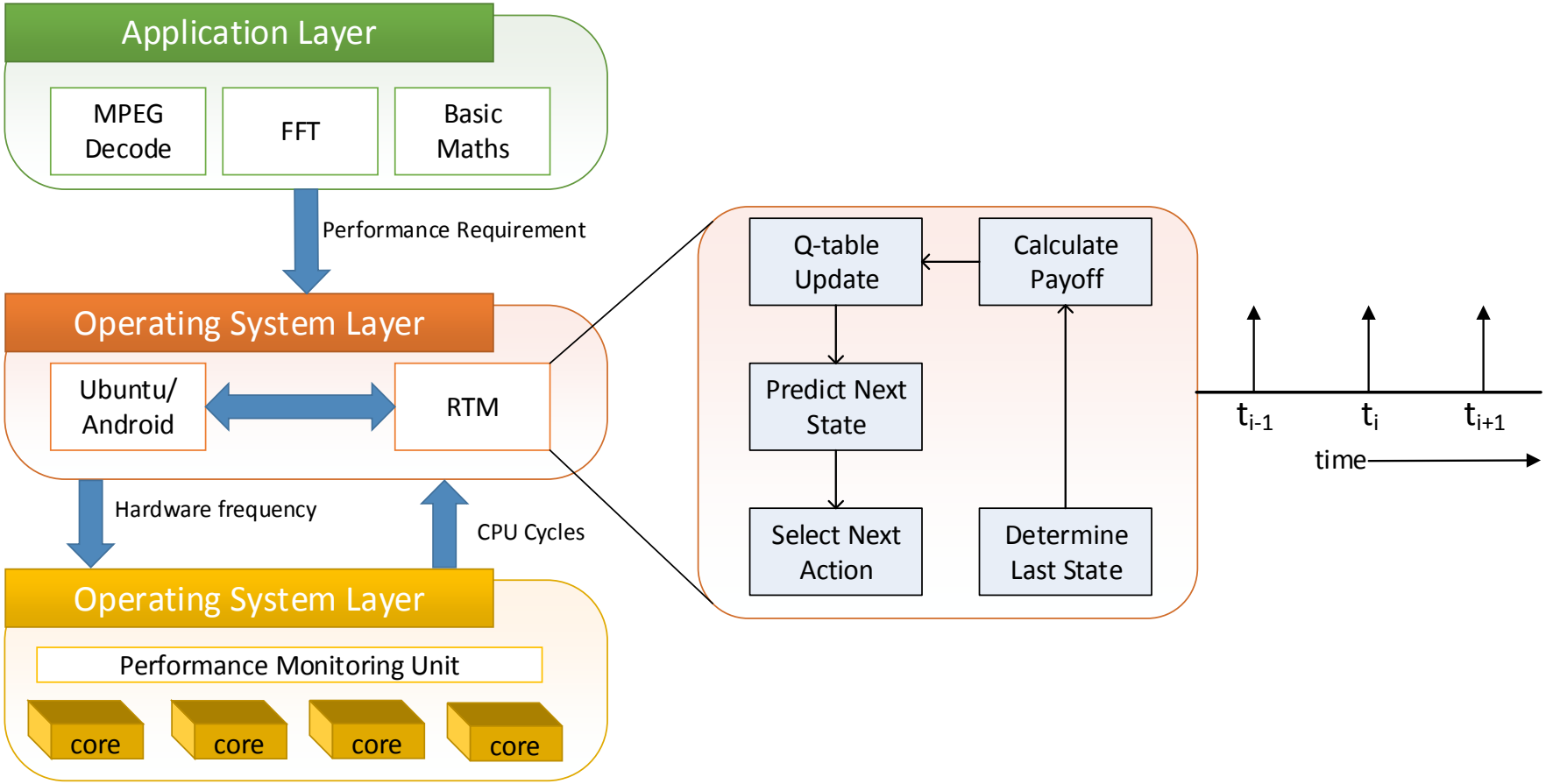
- **Dynamic voltage and frequency scaling (video example)**



- **Optimization Algorithm = Simple PI controller or advanced Reinforcement Learning**

Run-time Energy Optimization

- **Typical Low-level Flow**

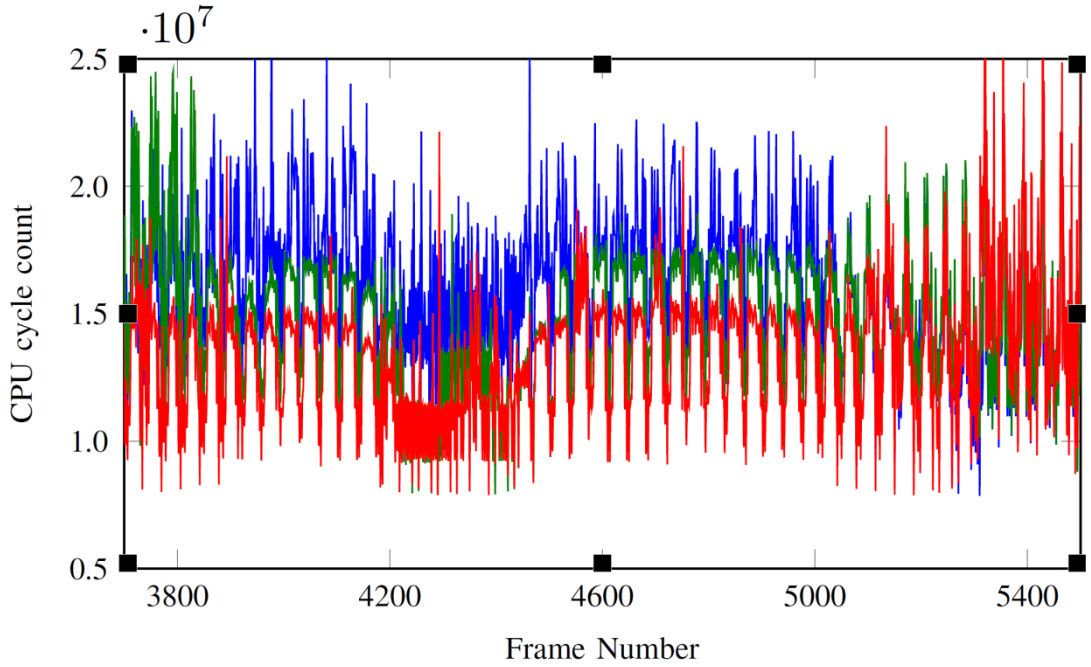


Workload Uncertainty

- Energy savings dependent on frame performance
  - CPU cycles measured from PMU
- Are the readings from PMU registers accurate?
  - Different energy results at different iteration
- PMU registers recorded at multiple iteration to investigate this

Workload Uncertainty

- **Experiment on video decoder, recording CPU cycle counter for three consecutive runs of same video**



60% variation
For some frames

- **Observation: Green reading higher around 3800 frames, blue around 4200 and red around 5400 frames**

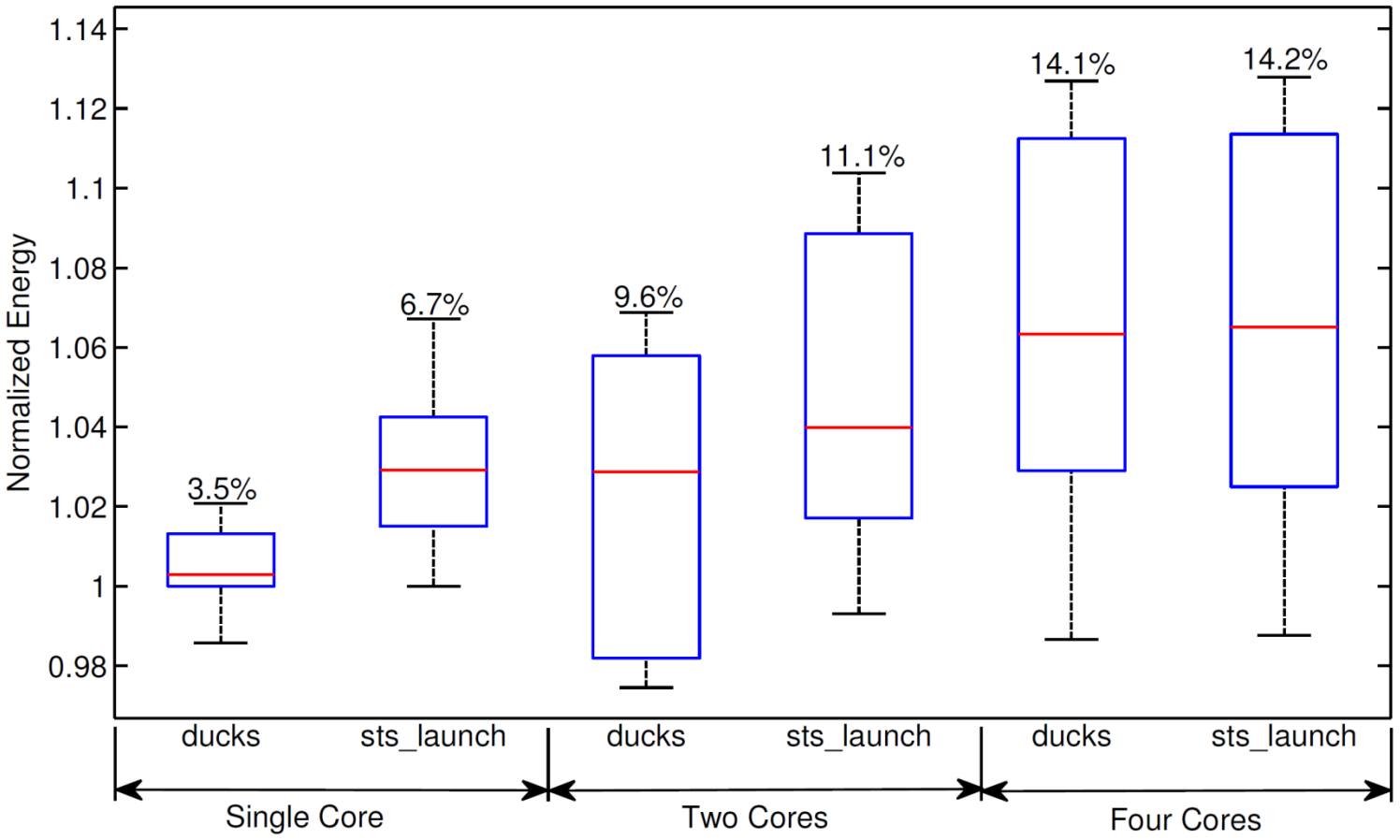
Workload Uncertainty – Impact

- **W = actual workload**
- **W‘ = observed workload**
- $W' = W + e$ **(uncertainty)**
- **Frequency proportional to workload**
- $\frac{f_{applied}}{f_{required}} = \frac{W'}{W} = 1 + \frac{e}{W} > 1$

- **Impact on energy**

Workload Uncertainty – Energy Impact

- **Depends on the number of cores**



Proposed Approach

- **Objective: Run-time energy optimization**

- **Question: How to use workload uncertainty at our benefit to improve energy?**

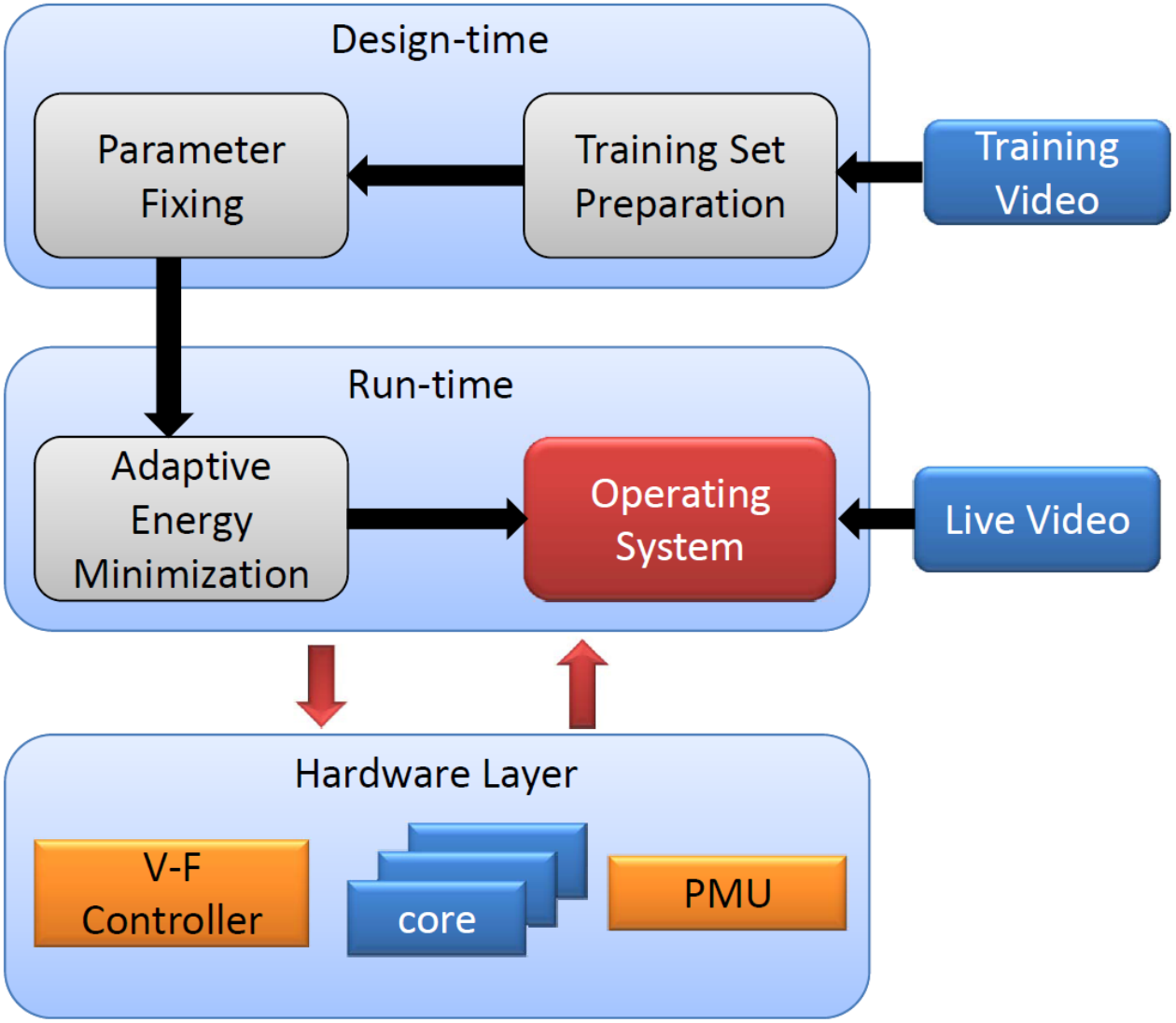
Proposed Approach – Foundation

- Hardware frequencies controlled by OS are discreet for any platform
- Do we need to make frequency proportional to workload and then discretizing?
- Can we have a classification approach?
  - Each frequency level is a class
  - Mapping of a workload to a class
  - Requirement: mapping for known workloads need to be known apriori
- Logistic regression for multiple levels – Multinomial Logistic Regression (MLR)

Proposed Approach – Hybrid Approach

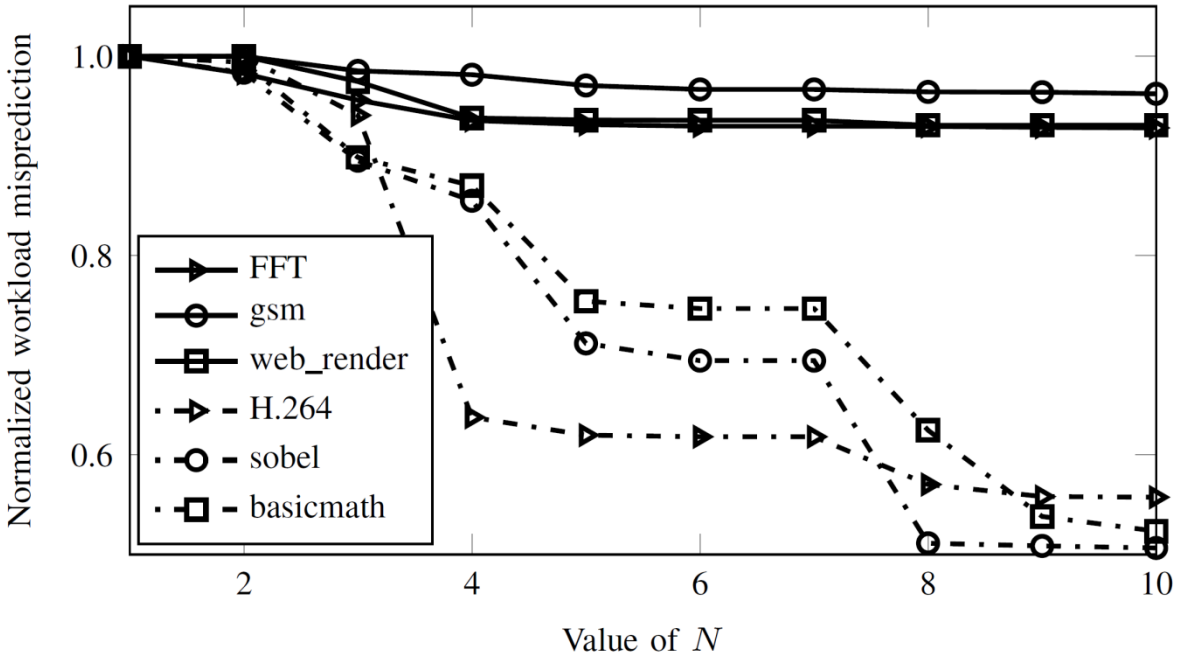
- MLR to known workload (workload and class known)
- Frame-based processing example
  - Inputs = Previous N frame data + current frame data
  - Output = Probability that current frame belongs to different classes
  - Select the class with the highest probability
  - Compare with known class and fix model parameters
- At run-time, apply trained model to unknown workloads
- Supervised learning

Run-time Energy Minimization



MLR Choice

- **Window of frames (for video GOP)**



- **Dynamic workload (H.264) requires a window of frames for accurate prediction**
- **Static workloads can be predicted using previous frame**

Contributions

- Video = I, P, B frames
- Previous frame workloads are not all independent
  - MLR is suitable
- MLR using just the previous frame has been solved. May not be accurate for dynamic workloads

- Contribution:
  - First approach to consider workload uncertainty in MLR and use it for energy optimization
  - DVFS overhead

Contributions

- **1: Generate MLR model with uncertainty**

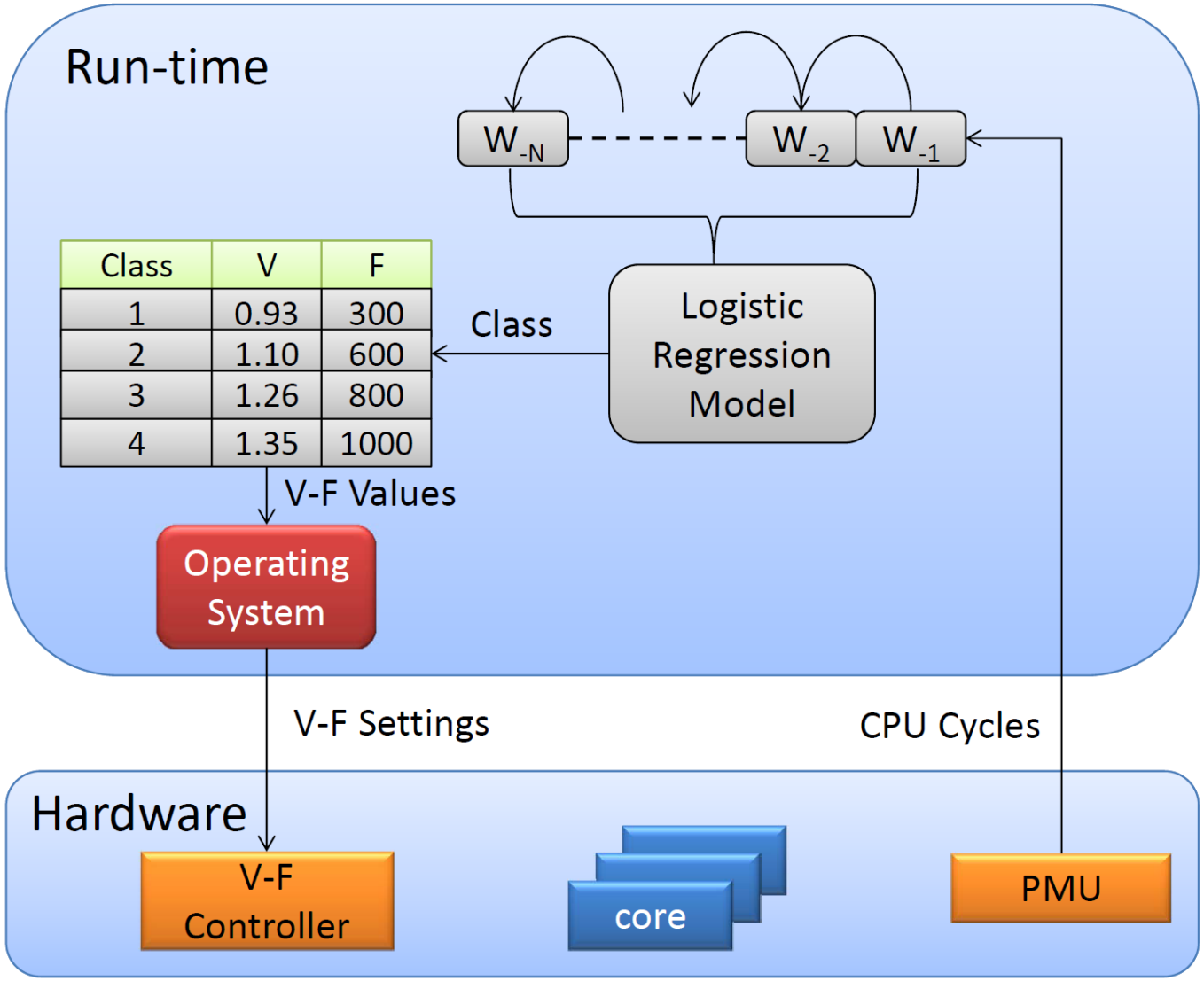
$$h_\theta(X) = \left[\left(\sum_{r=1}^{K} \gamma_{1,r} \cdot p_r\right) \quad \cdots \quad \left(\sum_{r=1}^{K} \gamma_{K-1,r} \cdot p_r\right)\right]^T$$

- **2: Parameter fitting**
  - **Maximum likelihood estimation**

$$\ell(\theta) = \ln\left(\mathcal{L}(\theta)\right) = \sum_{i=1}^{M}\sum_{l=1}^{K} \mathcal{I}(y^{(i)} = l) \cdot \ln\left(\frac{e^{\left(\theta^{(l)T} \cdot X^{(i)}\right)}}{\sum_{j=1}^{K} e^{\left(\theta^{(j)T} \cdot X^{(i)}\right)}}\right)$$

- **3: Run-time probability estimation**
- **4: Map probability to a class i.e., voltage-frequency pair considering voltage-frequency switching overhead**

Revisit the Overall Approach



Results – Energy Consumption

- **Comparisons**
  - **Linux Ondemand**
  - **Predictive [1]**
  - **Qlearning [4]**
  - **MLR [9]**

[1] Jung et al. “Continuous Frequency Adjustment Technique Based on Dynamic Workload Prediction” in VLSI Design 2008
[4] Shen et al. ““Achieving Autonomous Power Management Using Reinforcement Learning,” ACM TODAES, 2013.
[9] Cochran et al. “Identifying the Optimal Energy-Efficient Operating Points of Parallel Workloads,” in ICCAD, 2011.

Results – Energy Consumption

- **Comparisons**
  - **Linux Ondemand**
  - **Predictive [1]**
  - **Qlearning [4]**
  - **MLR [9]**

[1] Jung et al. “Continuous Frequency Adjustment Technique Based on Dynamic Workload Prediction” in VLSI Design 2008
[4] Shen et al. ““Achieving Autonomous Power Management Using Reinforcement Learning,” ACM TODAES, 2013.
[9] Cochran et al. “Identifying the Optimal Energy-Efficient Operating Points of Parallel Workloads,” in ICCAD, 2011.

Results – Energy Consumption

- Comparisons
  - Linux Ondemand
  - Predictive [1]
  - Qlearning [4]
  - MLR [9]
  - Proposed (no DO no WN) ➔ w/o DVFS Overhead Optimization, w/o workload uncertainty
  - Proposed (DO, no WN) ➔ w/o workload uncertainty
  - Proposed with all knobs enabled

Results – Energy Consumption

TABLE I: Energy consumption (Joules)

Videos	Ondemand [11]	Predictive [1]	Q-Learning [4]	MLR [9]	Proposed		
					no DO, no WN	DO, no WN	DO, WN
FFT	32.7	29.0	27.6	20.2	19.8	19.6	16.9
gsm	18.8	14.3	15.5	10.8	10.6	10.5	8.8
web_render	26.6	26.6	24.5	21.5	21.5	21.1	20.2
H.264	23.7	23.7	20.2	18.5	18.2	17.9	15.4
sobel	69.3	61.4	63.2	41.4	30.7	29.7	28.3
basicmath	82.6	82.6	60.2	42.4	40.9	40.0	37.3

Proposed (no DO, no WN) equally good as MLR. The improvement is because of window of frames.

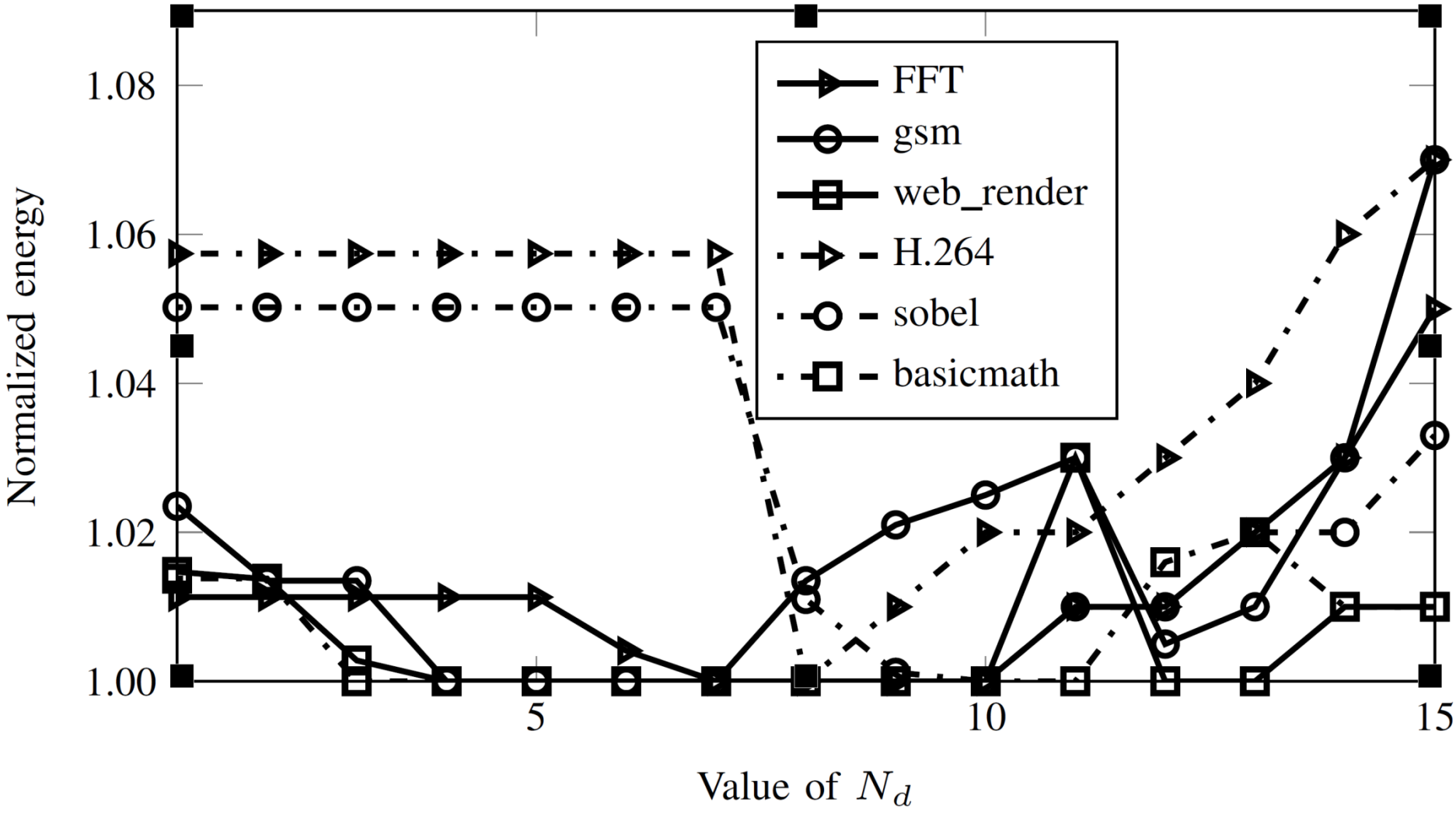
Proposed (DO, no WN) better, signifying the impact of switching overhead every frame as opposed to every window of frames

Proposed (DO, WN) even better, due to correct estimation of uncertainty

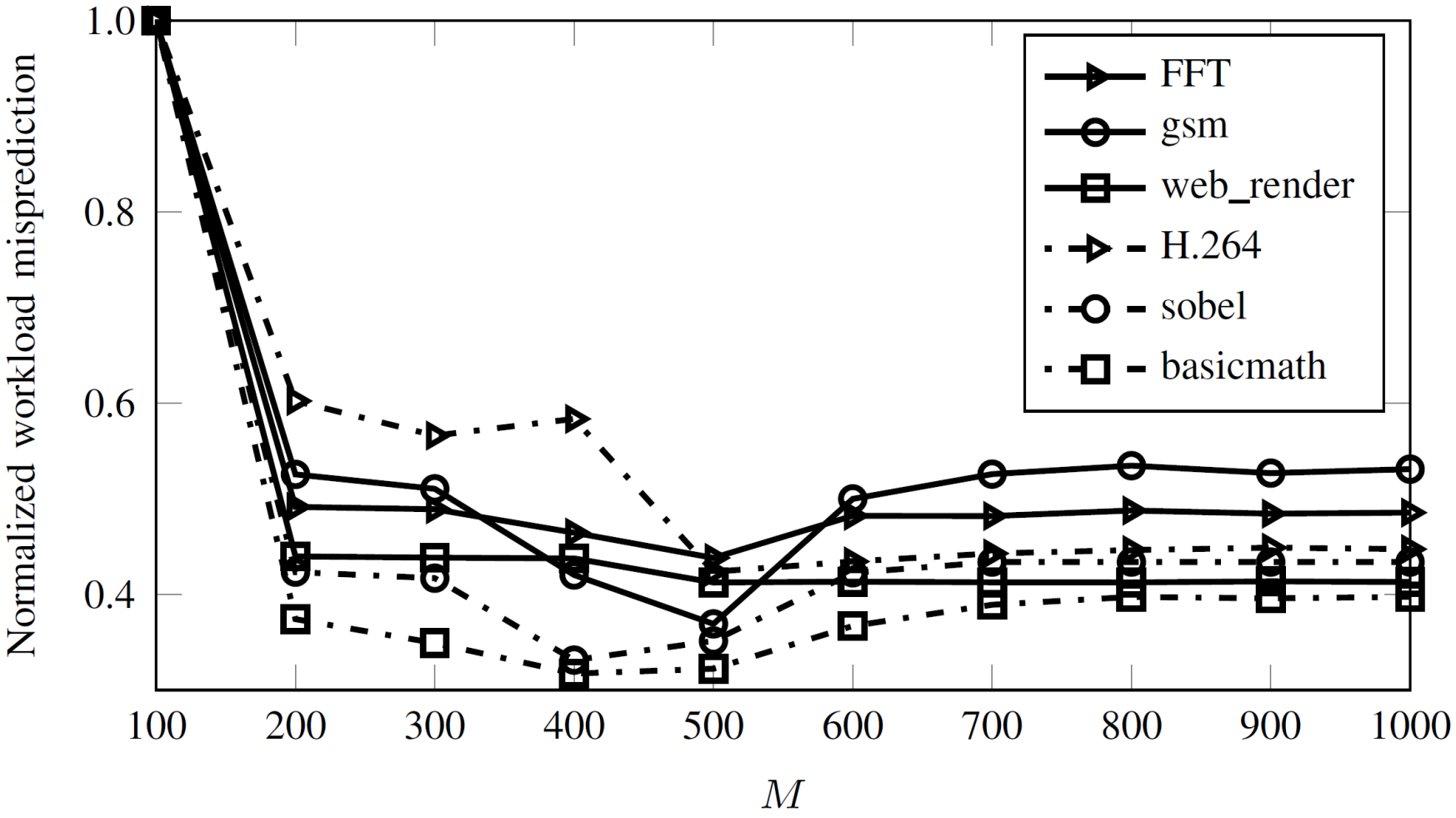
Results – Energy Consumption summary

- **Proposed approach 20% lower energy (average)**
  - **11% due to workload uncertainty**
  - **8% due to frame history based prediction**
  - **Rest due to DVFS overhead**

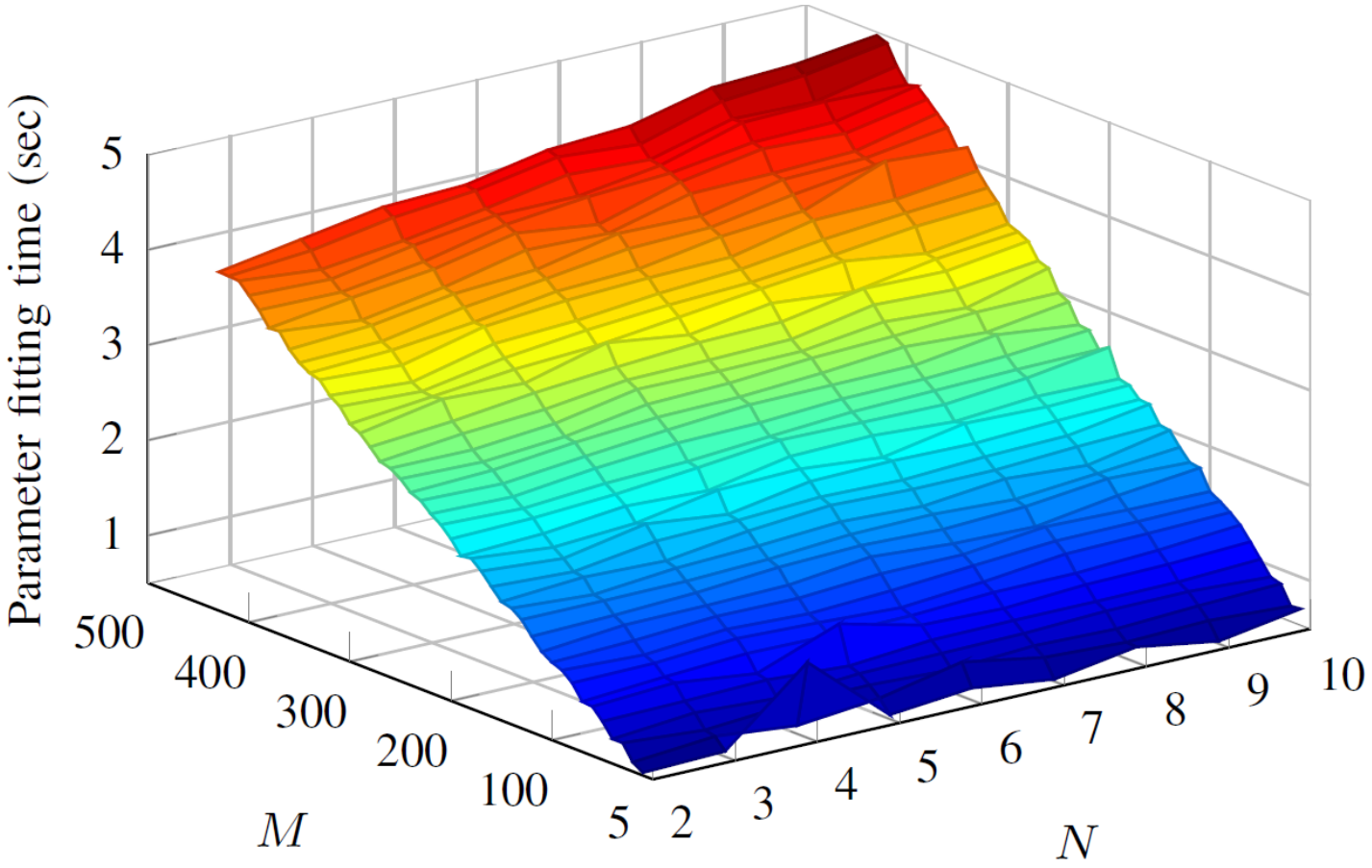
Results – Parameters (DVFS Window Size)



Results – Parameters (Training set size)



Results – Parameters (Training time)



Conclusions

- **Hybrid approach for energy minimization**
- **Multinomial logistic regression**
- **Incorporate workload uncertainty**
- **Incorporate DVFS switching overhead**
- **20% improvement over state-of-the-art approaches**

Thank You!!

www.prime-project.org
www.anupdas.com (a.k.das@soton.ac.uk)

DATE Booth: EP1