Workload Prediction and Dynamic Voltage Scaling for MPEG Decoding

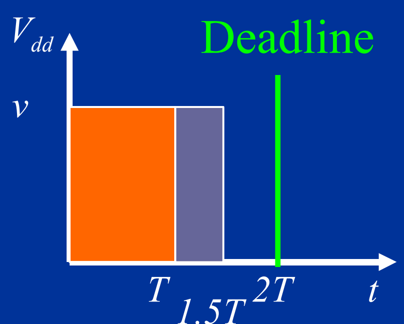
Ying Tan, Parth Malani, Qinru Qiu, Qing Wu
Dept. of Electrical & Computer Engineering
State University of New York at Binghamton

Outline

- Introduction
- Background on MPEG decoding
- Proposed workload prediction and DVFS techniques for software MPEG decoding
- Experimental results
- Conclusions

Dynamic Voltage/Frequency Scaling

- Using DVFS with buffer reduces the energy even more
  - Borrow or steal processing time from adjacent tasks
  - But latency and hardware complexity also increases



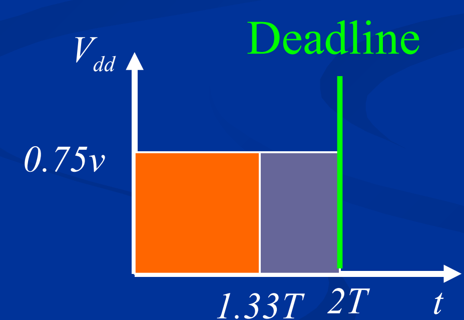
Without DVFS

$E_1 = C_L * V^2 * f * (1.5T)$

Without Buffer

$E = C_L * (f * V^2 * T + f/2 * V^2/4 * T)$

$= 0.75E_1$



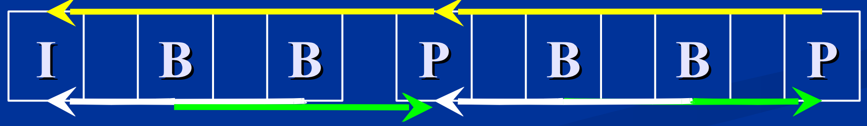
With Buffer

$E = C_L * 0.75f * (0.75)^2 V * 2T$

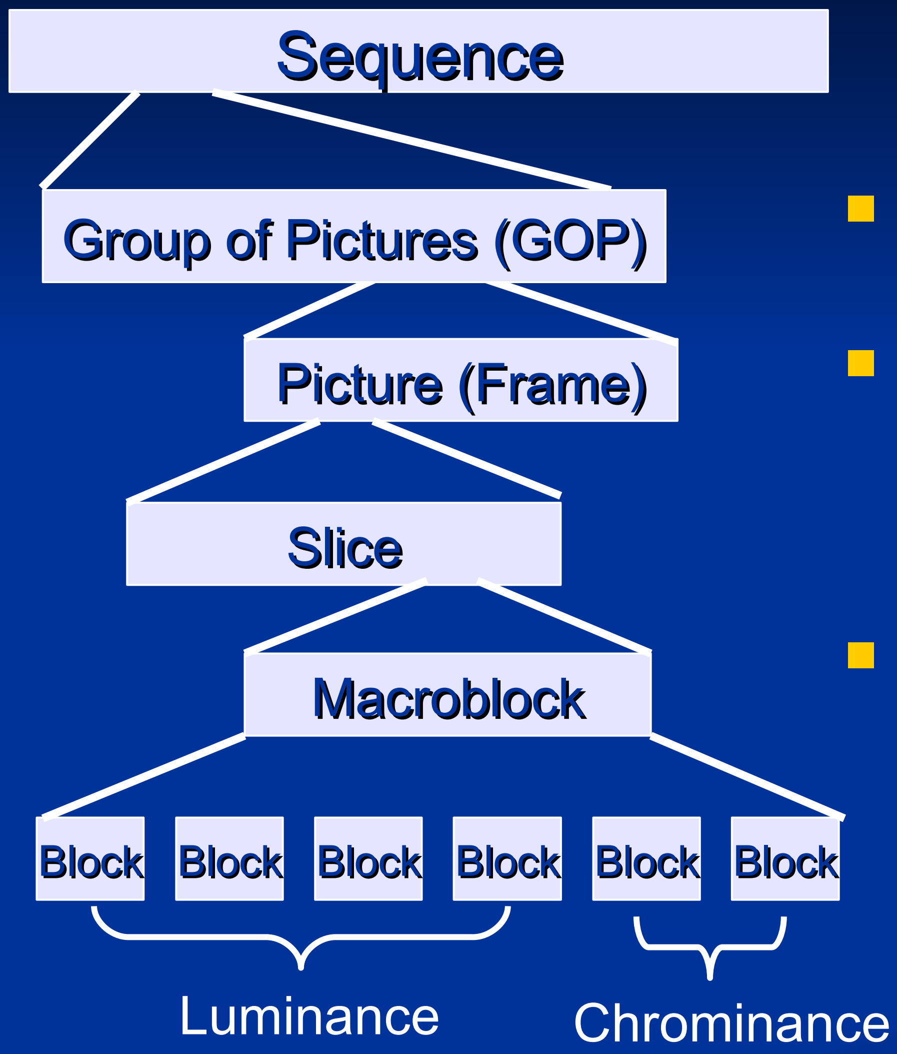
$\approx 0.56E_1$

MPEG-Frame Types

- Video stream: a sequence of still images (frames)
- **I**-frames (*intra-coded* frames) do not depend on any other frame
- **P**-frames (*predictive coded* frames) are encoded using past I or P frame as a reference
- **B**-frames (*bi-directionally predictive coded* frames) use both past and future I or P frames as references



MPEG-Layered Structure



- A GOP is an independently decodable unit that begins with an I-frame
- A macroblock is a 16X16 pixel area image
- A block is a 8X8 pixel area of image which carries only luminance or chrominance information
- Macroblocks can be divided into four types

MB / frame	I	P	B	Bi
I	X			
P	X	X		
B	X	X	X	X

Workload in MPEG Decoding

- The number of instructions to perform one IDCT or motion compensation is almost a constant for a given processor
  - Only need to count the number of IDCT and motion compensation

	I	P	B	Bi
IDCT only	X	X	X	X
IDCT+FW		X		
FW only		X		
IDCT+BW			X	
BW only			X	
IDCT+Bi				X
Bi only				X
Skipped		X	X	X

- IDCT and motion estimation is done at block level
  - Blocks are divided into 8 different types
  - Decoding time of each type of block is assumed to be a constant

Workload Prediction

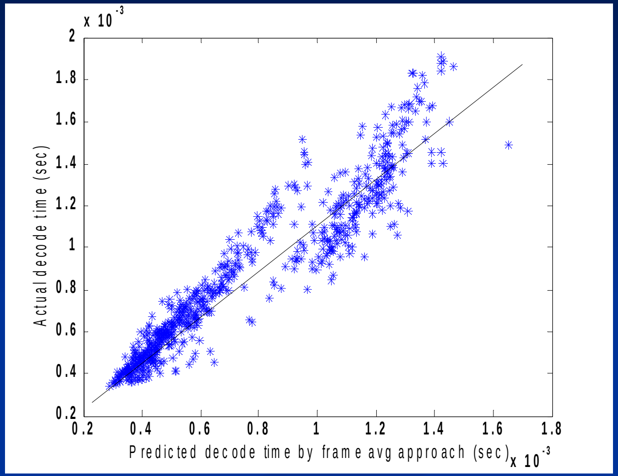
- Our workload predictor is a linear model
  - Variables M1~M8 represent the number of 8 different types of blocks
    - The information could be obtained from the macroblock header
  - Variable M9 represents the frame size
  - Coefficients are obtained using linear regression analysis

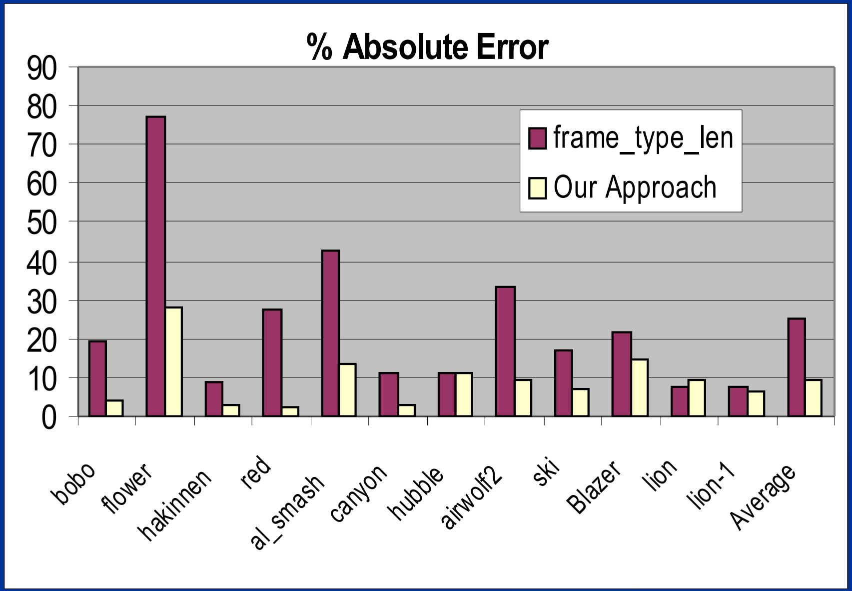
$$frame\_decode\_time = w_0 + \sum_{1 \le i \le 9} w_i \cdot M_i$$

Comparison with Existing Predictor

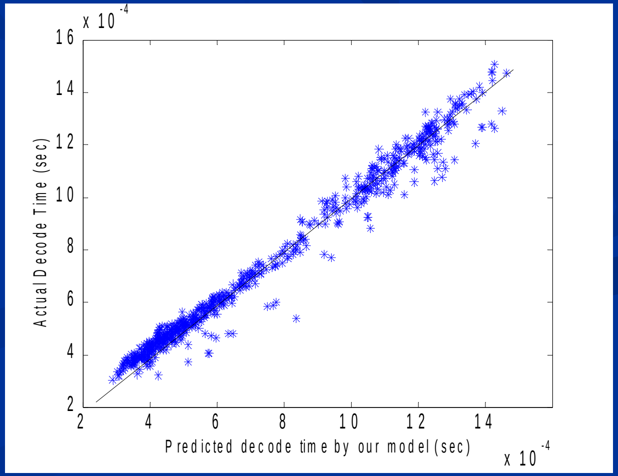
- Berkeley MPEG decoder running on Pentium IV 2.6GHz processor
- Frame_Type_Len: moving average of previous decoding time combined with frame size

Frame_Type_Len





Our Approach



Optimal DVFS

- Assumptions
  - Continuous frequency/voltage scaling
  - Negligible switching cost
  - Input and display at a constant rate whose period is T
- The optimal DVFS is to decode every frame continuously without any pause in nT time at a constant frequency and voltage, where n is the total number of frames in a video stream
  - Does not consider arrival time and display deadline
    - These constraints can be met by adding input/output buffers and increasing the latency
  - Must have the workload information of the entire stream
  - Lowest energy, however, highest buffer requirement

GOP-Optimal DVFS

- Buffers all the frames in a GOP and decodes the entire GOP using a constant voltage
  - On-line heuristic of Optimal DVFS
  - Does not consider the frame incoming time and display deadline
    - In the worst case the input buffer needs to be 2 GOP long

Global Grouping

- Divide the time into *n* intervals $D_1 \sim D_n$ based on display deadline
- Consecutive intervals $(D_i, D_{i+1}) \sim (D_{k-1}, D_k)$ will be grouped together if we can find a constant voltage/frequency such that the processor can decode frame i~k continuously before their deadline without pausing

Global Grouping

- The processor is running at a steady speed within the time slots in a group;
- The complexity of global grouping is $O(n^2)$
- The global grouping is an off-line algorithm since it requires the workload information for the entire stream
  - More suitable for the movie clips that are played repeatedly
- It has minimal energy dissipation while meeting the deadline if all the frames are available at the beginning

Dynamic Grouping

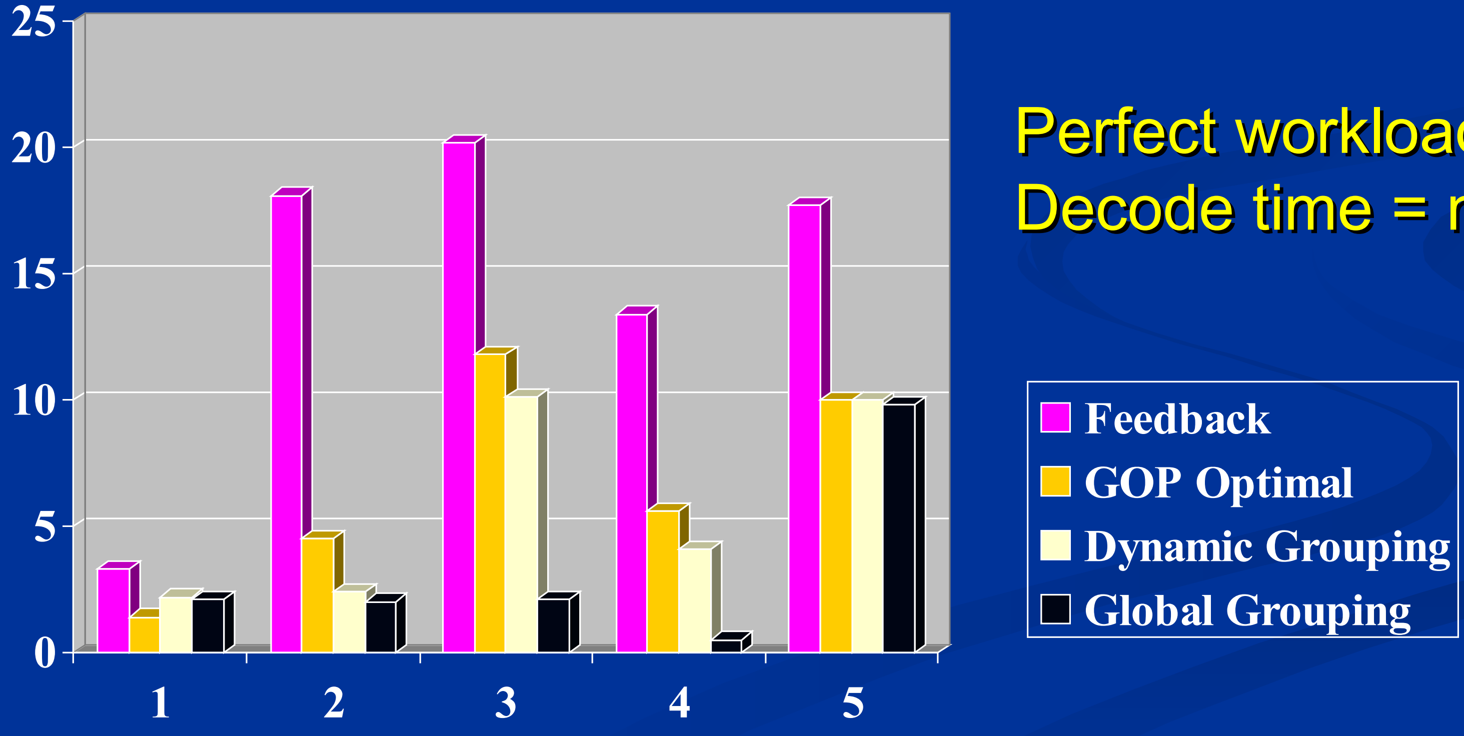
- Buffers the input frames up to a certain *window* size at the beginning, applies the global_grouping within the window
- When a new frame with workload *x* comes in, (*avg_load* is the average workload for the last group in current window)
  - if x < avg_load, make it an individual group
  - if x = avg_load, merge it into the last group
  - if x > avg_load, merge it into the last group i, and recalculate the average workload for each group
- The dynamic grouping is an on-line heuristic of global grouping. It gives better trade off between energy and buffer size

Characteristics of MPEG Clips

MPEG Clips		*Frame Type*	*# of Frames*	*GOP Size*
Name	*Index*			
hakkinen	1	I,P,B	799	12
bobo	2	I,P,B	679	90
ski	3	I,P,B	1513	15
blazer	4	I,P,B	2998	12
wg	5	I,P	130	6

Experimental Results – Energy

- DVFS using feedback control
  - A controller is used to adjusts the decoder's speed to keep a constant occupancy of the buffer between the decoder and the display



Perfect workload prediction.
Decode time = nT

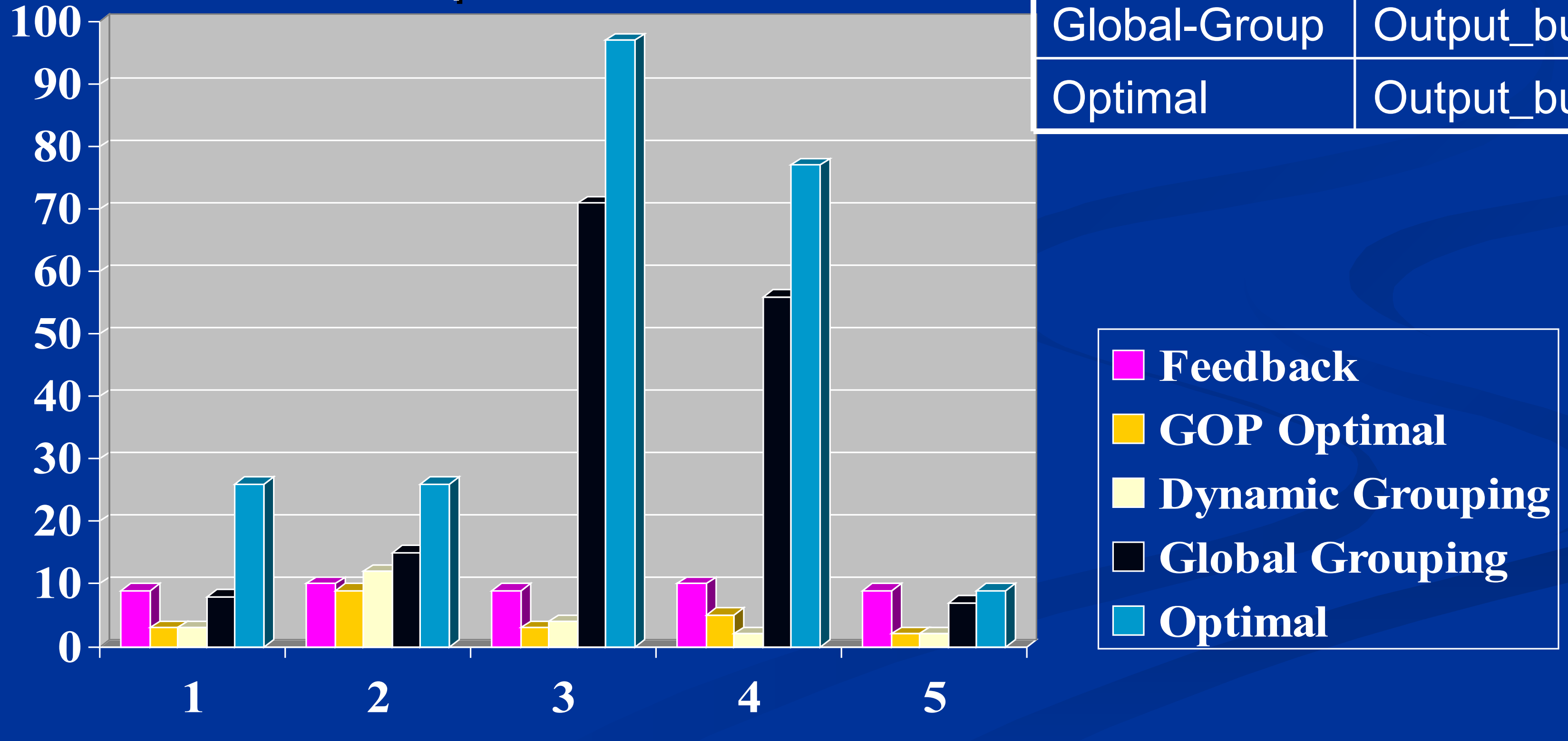
Experimental Results - Buffers

Perfect workload prediction.
Decode time = nT

Input Buffer

DVFS	Input Buffer
GOP-Opt	2GOP
Dyn-Group	1GOP
Global-Group	Output_buffer±1
Optimal	Output_buffer±1

Output Buffer



Summary

- The proposed workload prediction model utilizes the block level statistics of each MPEG frame and gives highly accurate prediction results
- Proposed DVFS techniques give good energy reduction, less buffer usage and work robustly with our predictor