A Case for Dynamic Pipeline Scaling

Prakash Ramrakhyani,
Jinson Koppanalil*, Sameer Desai, Anu Vaidyanathan, Eric Rotenberg

Center for Embedded Systems Research (CESR)
Department of Electrical & Computer Engineering
North Carolina State University
www.tinker.ncsu.edu/ericro

*Arm Incorporated, Austin, TX

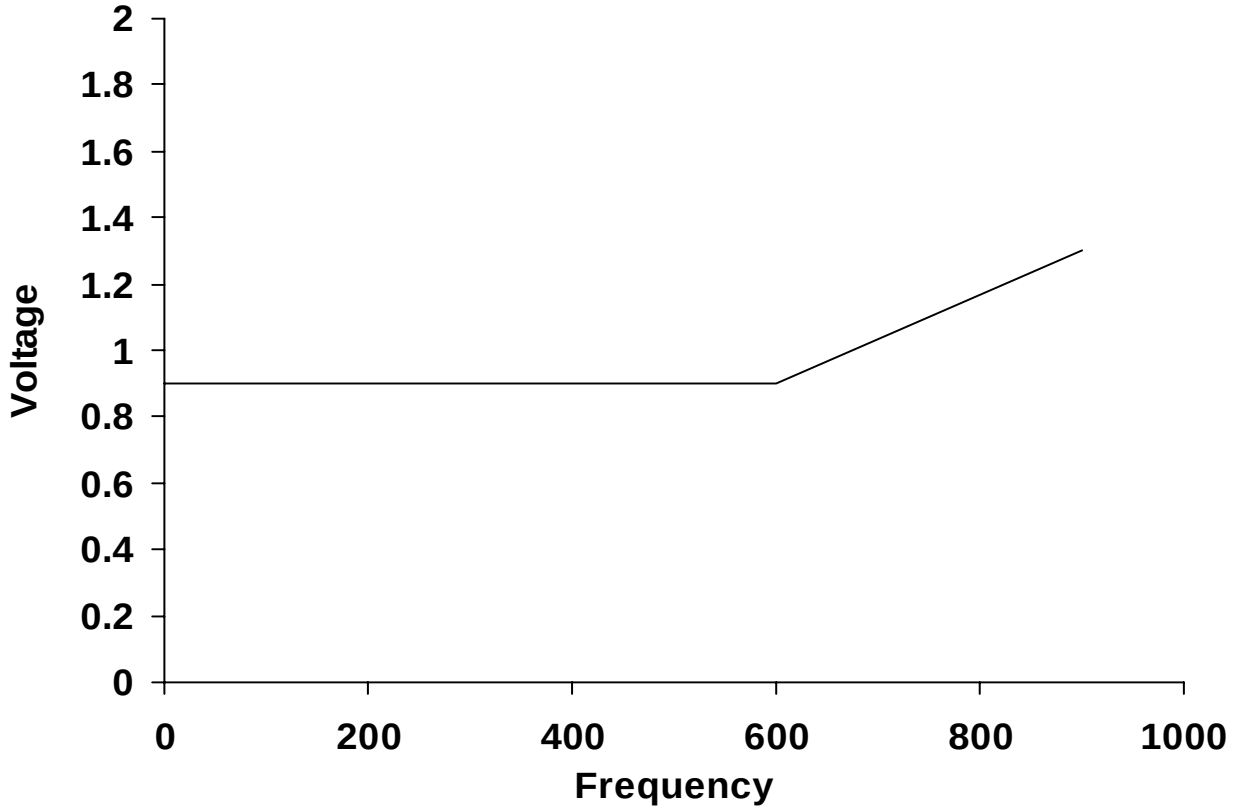
Traditional Energy Management

- Save energy by lowering frequency when peak performance not needed
  - Extended clock period means we can increase logic delay

$$\text{delay} \propto \frac{1}{V}$$

  - Hence can reduce voltage
  - $E \propto V^2$, reducing voltage saves energy
    - Dynamic Voltage Scaling (DVS)
- But, how low can the voltage get?

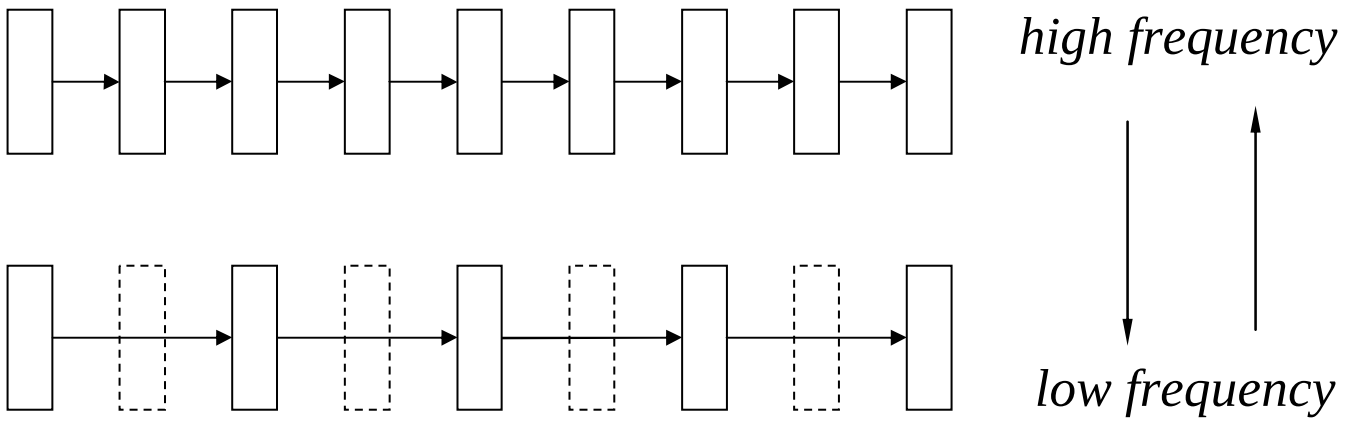
Limitations of DVS



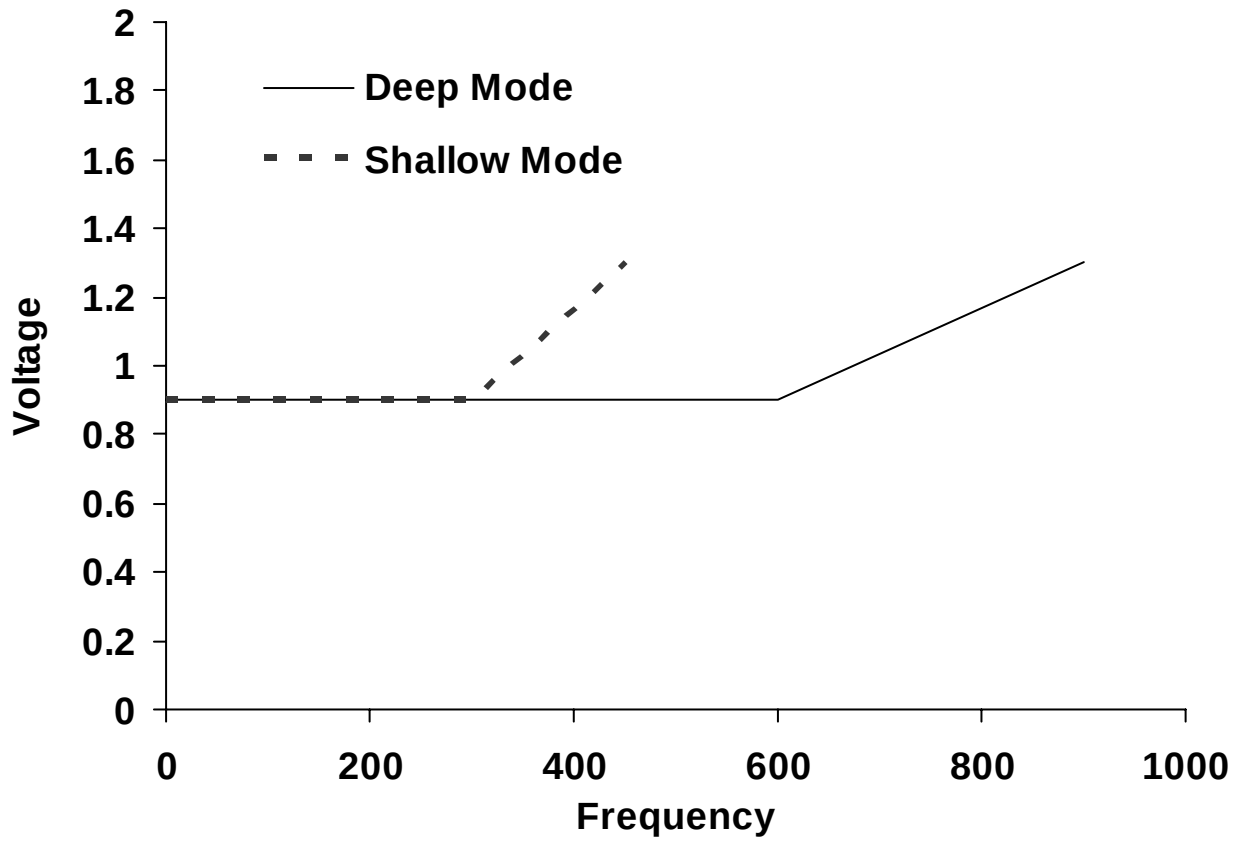
- Lower voltage bound limits the useful frequency range of DVS

Dynamic Pipeline Scaling

- Alternative way to exploit lower frequency when V_{low} is reached
  - Merge adjacent pipeline stages



Modified Voltage-Frequency Characteristic



Why DPS Works

- Energy also depends on IPC

$$\text{Energy} \propto f \cdot V^2 \cdot t$$

$$\text{Energy} \propto f \cdot V^2 \cdot \left(\frac{\#\,\text{instr.}}{f \cdot \text{IPC}} \right)$$

$$\text{Energy} \propto \frac{V^2}{\text{IPC}}$$

- Deep pipeline has lower IPC than shallow pipeline
  - Longer data dependence stalls
  - Minimum misprediction penalty is twice that of shallow pipeline

Energy Differences between Deep and Shallow Modes

- Deep mode consumes more $useless\ energy$ than shallow mode:
  - More data stall cycles
  - More cycles spent executing down the wrong path
- We do not model:
  - Clock gating
    - Turn off unused units, reduce useless energy due to stalls
  - Fetch gating
    - Stop fetching on an unconfident branch
    - Reduce useless energy due to wrong-path instructions

Outline

- ✓ Introduction
  - Limited frequency range of DVS
  - DPS extends the frequency range
- Voltage - frequency characteristics
- Pipeline description
- Energy savings
- Summary
- Future work

Voltage-Frequency Characteristic

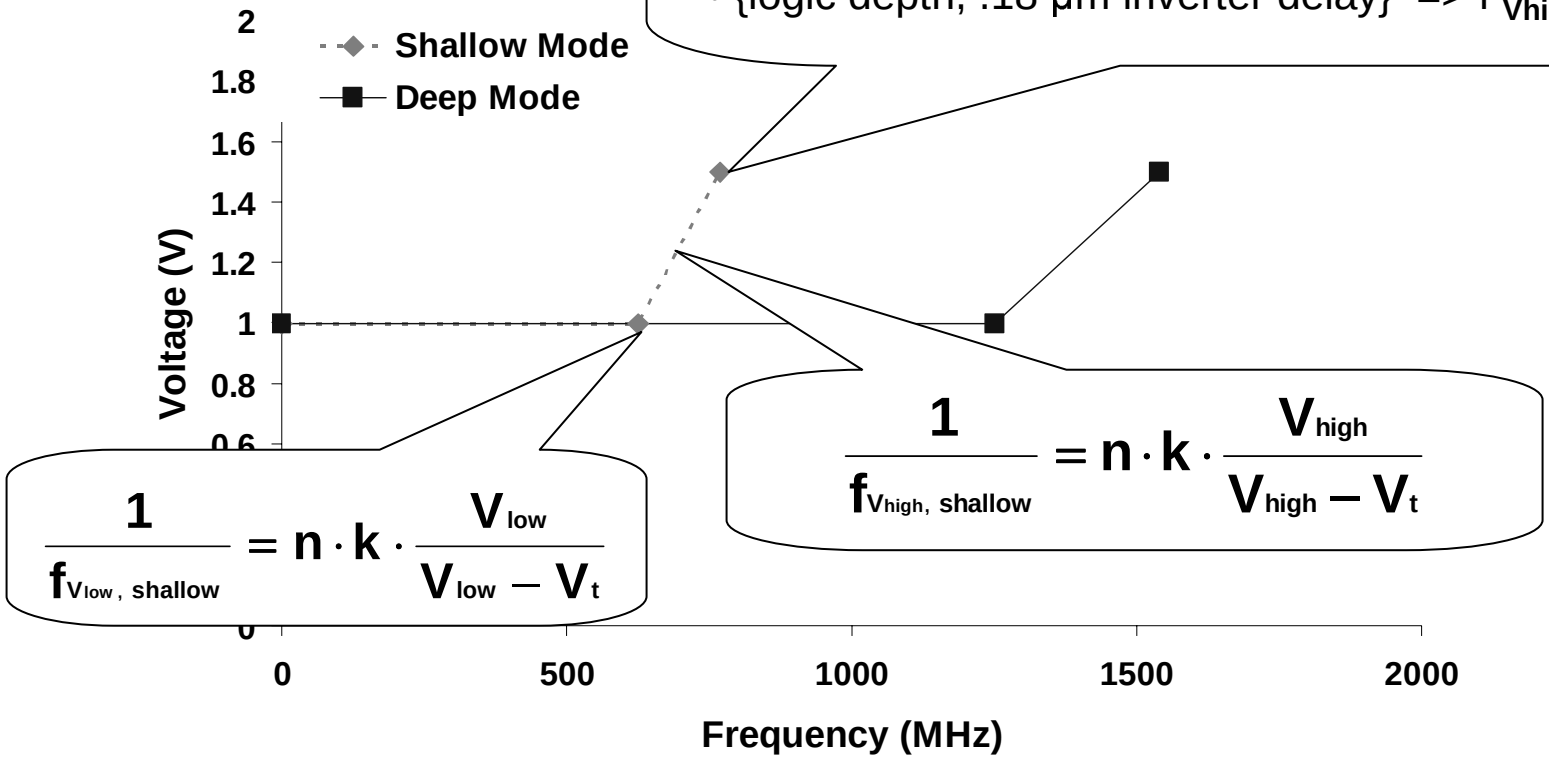
- Projected a V-f characteristic for an alpha-like processor
  - Alpha is similar to our shallow pipeline used to generate IPC's
- Corroborated our numbers with recent work, and real processors

Voltage-Frequency Characteristic

Voltage parameter	0.18µ		0.13µ	
	literature	TM5400	literature	TM5800
V_{high}	**1.5 V**	1.6 V	1.2 V	**1.3 V**
V_{low}	**1.0 V**	1.2 V	1.0 V	**0.9 V**
V_t	0.4 V	unspecified	0.3 V	unspecified

Voltage-Frequency Characteristic

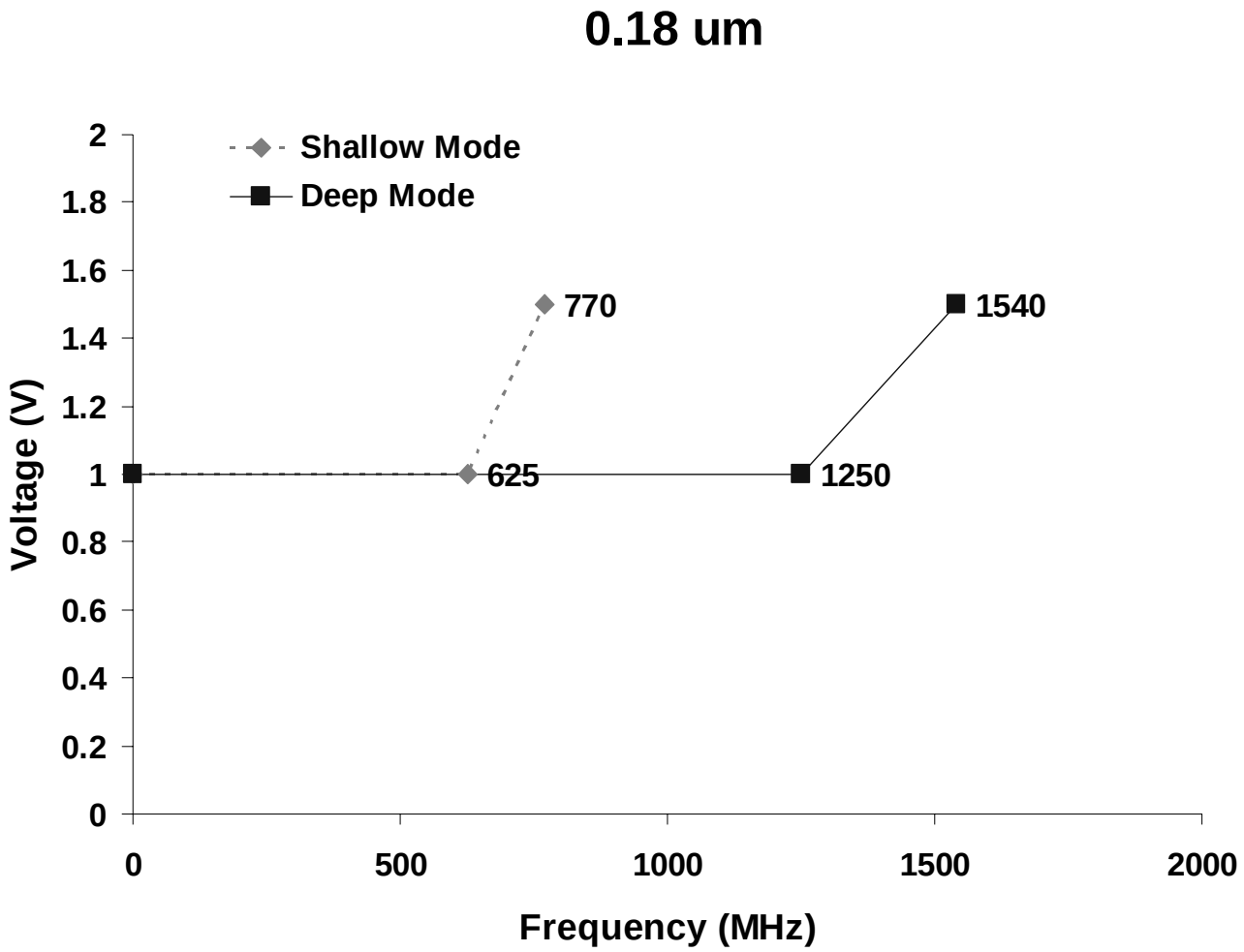
- Logic depth (n) of 1 pipe stage of Alpha 21264 {.35µm inverter delay, 400 MHz} => 20 FO4 inv.
- {logic depth, .18 µm inverter delay} => $f_{Vhigh,\ shallow}$



$$\frac{1}{f_{V_{high},\ shallow}} = n \cdot k \cdot \frac{V_{high}}{V_{high} - V_t}$$

$$\frac{1}{f_{V_{low},\ shallow}} = n \cdot k \cdot \frac{V_{low}}{V_{low} - V_t}$$

Voltage-Frequency Characteristic



Outline

- ✓ Introduction
- ✓ Voltage - frequency characteristics
- Pipeline description
- Energy savings
- Summary
- Future work

Deep Pipeline Mode

simple instructions (most integer ALU instructions)

IF1	IF2	ID1	ID2	W	S	RR1	RR2	EX1	EX2	WB1	WB2	RE1	RE2

complex instructions (integer multiply/divide, floating point)

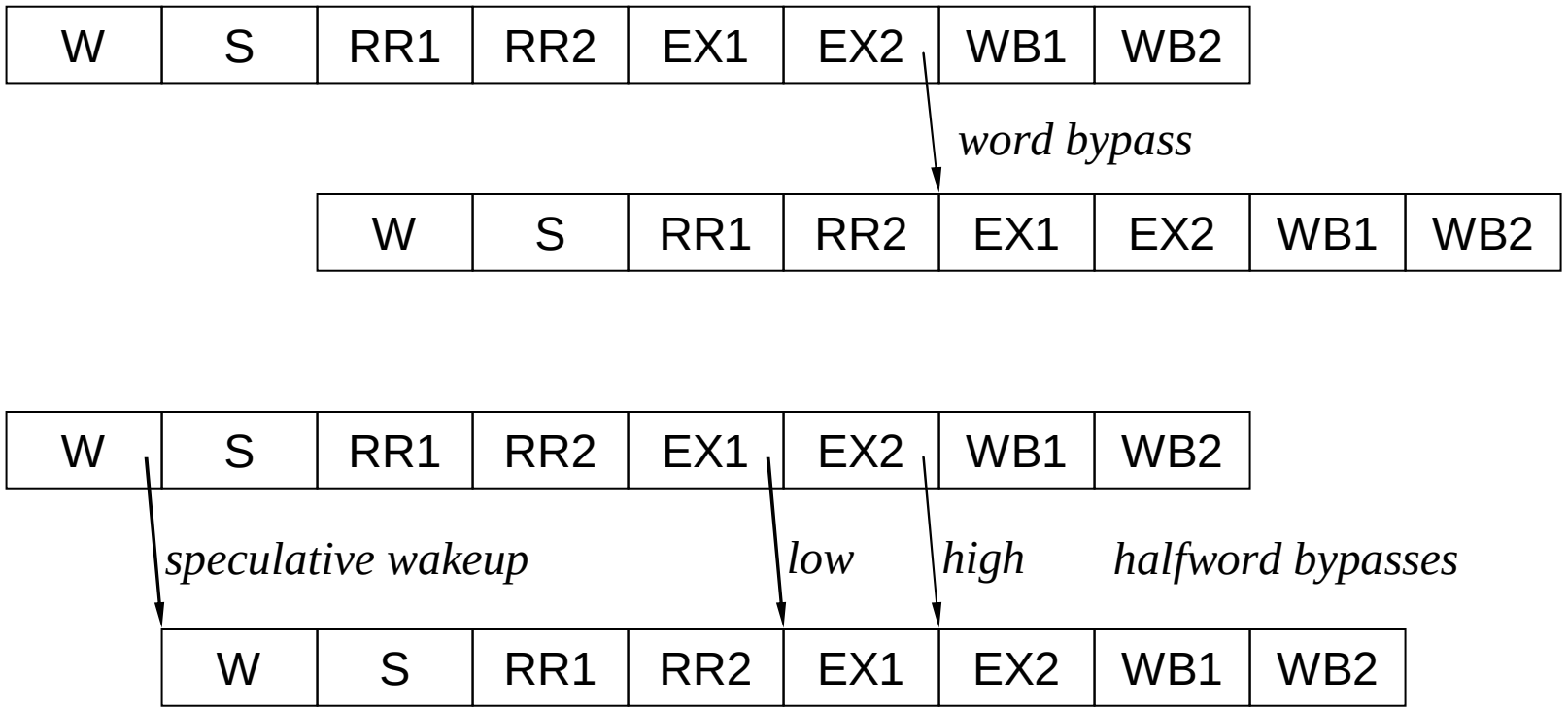
IF1	IF2	ID1	ID2	W	S	RR1	RR2	EX	...	WB1	WB2	RE1	RE2

loads/stores

IF1	IF2	ID1	ID2	W	S	RR1	RR2	A1	A2/M1	M2	WB1	WB2	RE1	RE2

Minimizing Data Dependence Stalls

- Halfword bypassing and speculative wakeup



Shallow Pipeline Mode

simple instructions (most integer ALU instructions)

IF	ID	IS	RR	EX	WB	RE

complex instructions (integer multiply/divide, floating point)

IF	ID	IS	RR	EX	...	WB	RE

loads/stores

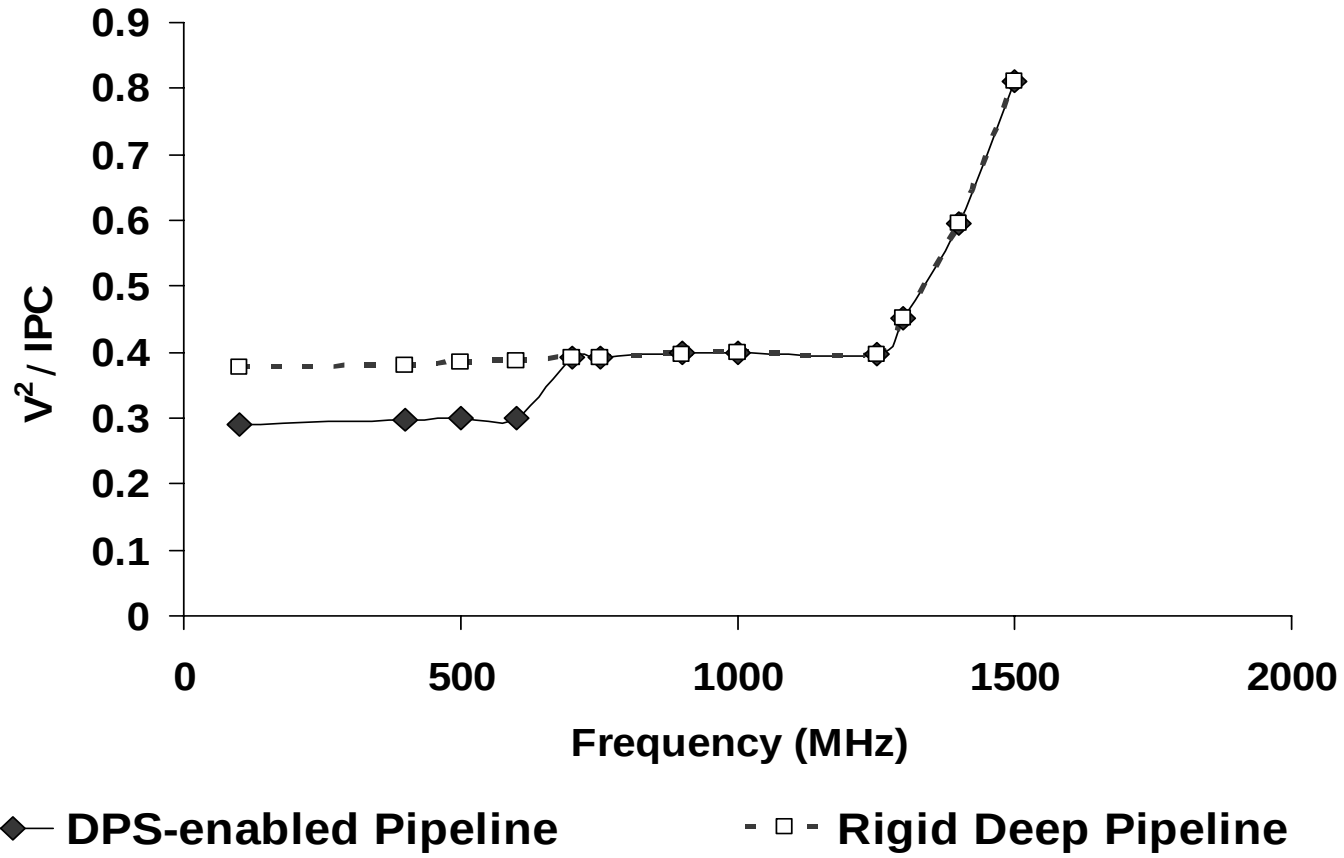
IF	ID	IS	RR	A	M	WB	RE

Simulation Environment

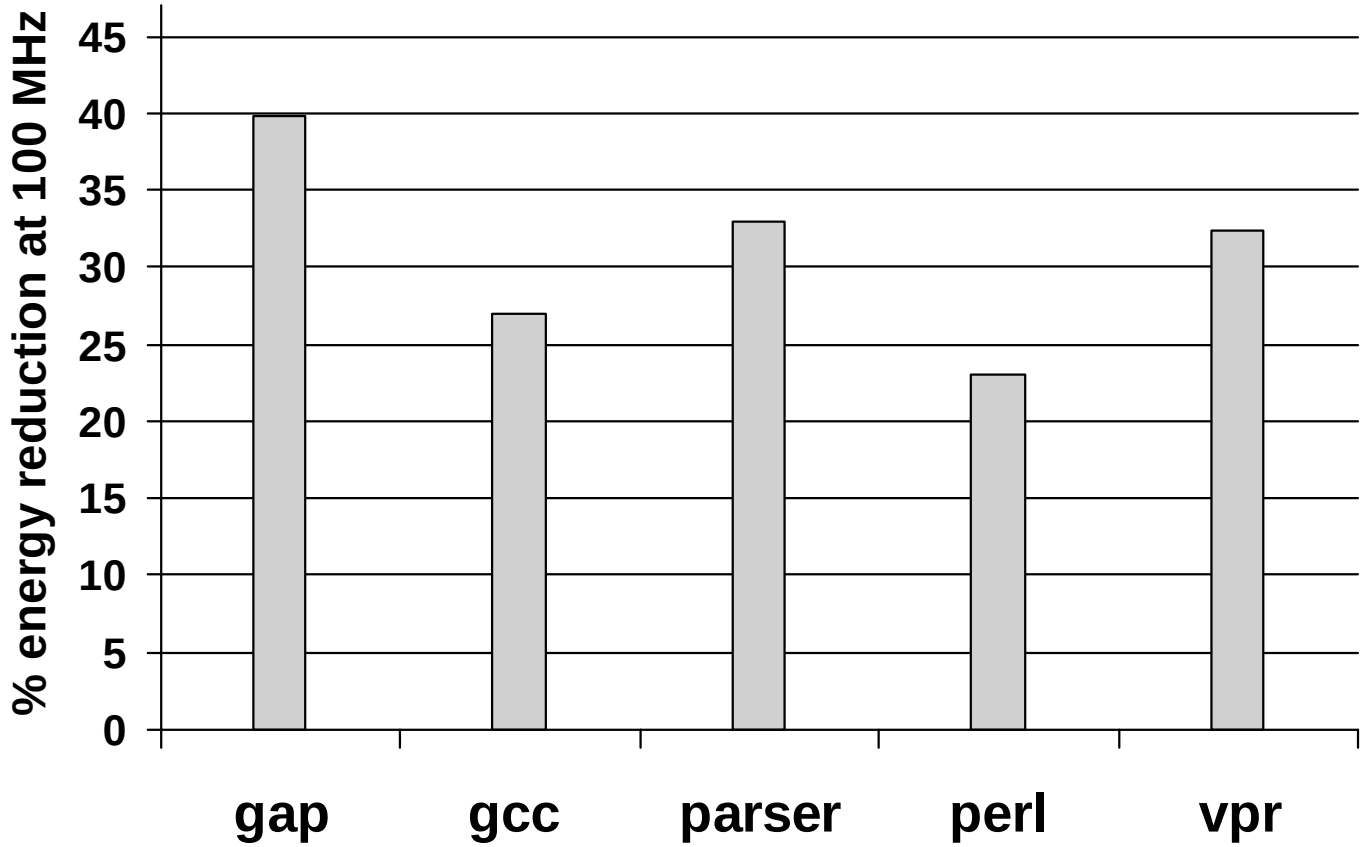
- A detailed cycle-accurate simulator
  - 8-way superscalar with 256-entry ROB
  - 64 K-entry gshare predictor & unlimited RAS
  - 32 KB L1 instruction and data caches
  - 512 KB unified L2 cache with 8 ns hit latency
  - Memory latency is 80 ns
- Energy Metric : $\frac{V^2}{IPC}$

Results

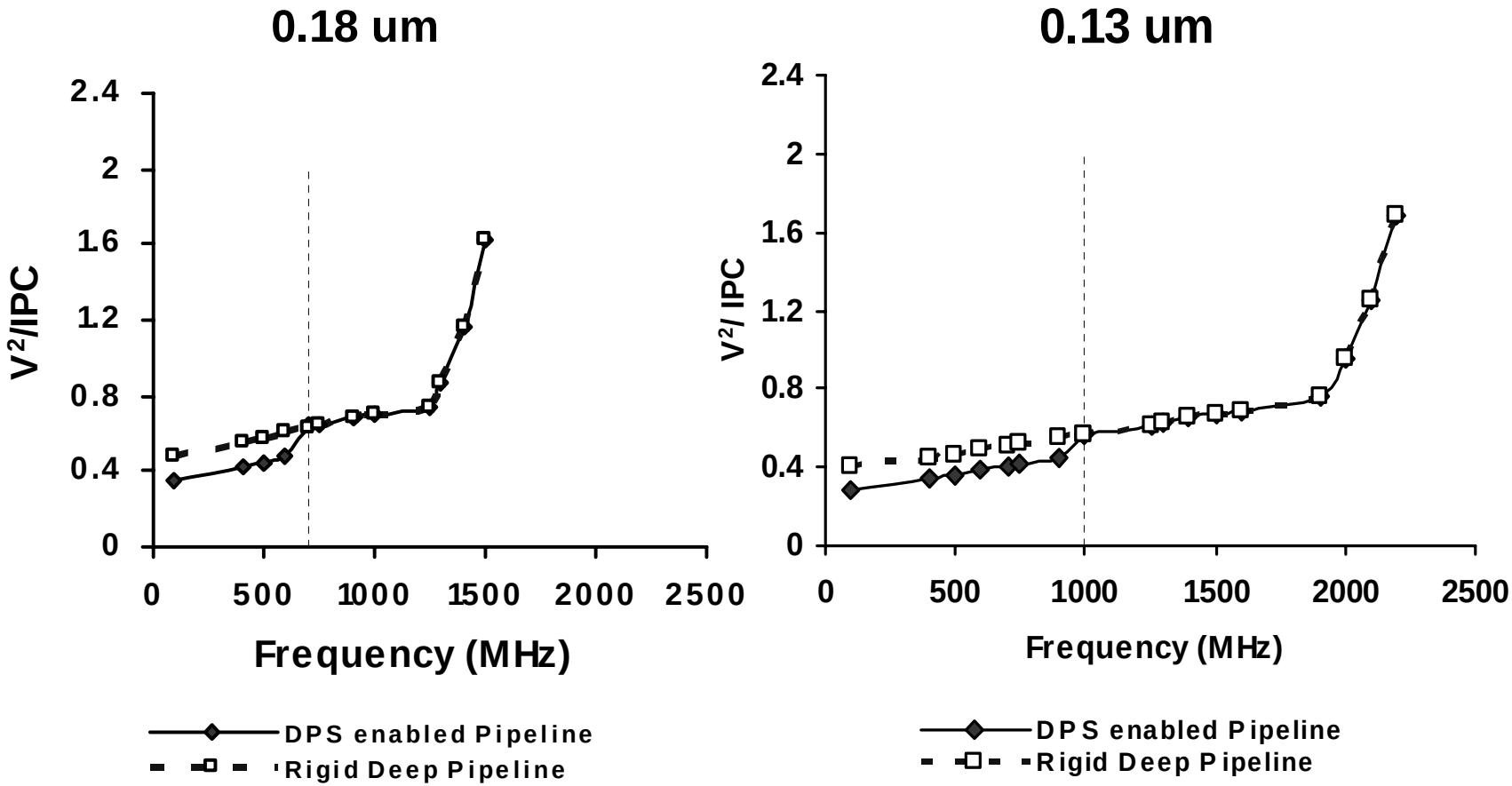
Benchmark: Perl



Results


% energy reduction at 100 MHz
45
40
35
30
25
20
15
10
5
0
gap
gcc
parser
perl
vpr

Effect of Technology Scaling



Summary

- DVS has a limited frequency range
- DPS: A technique to extend this range
  - Energy depends on IPC as well
  - Merge pipeline stages at frequencies below DVS range
  - Shallow pipeline has better IPC, hence lower energy
- 23-40% energy savings due to shallow mode

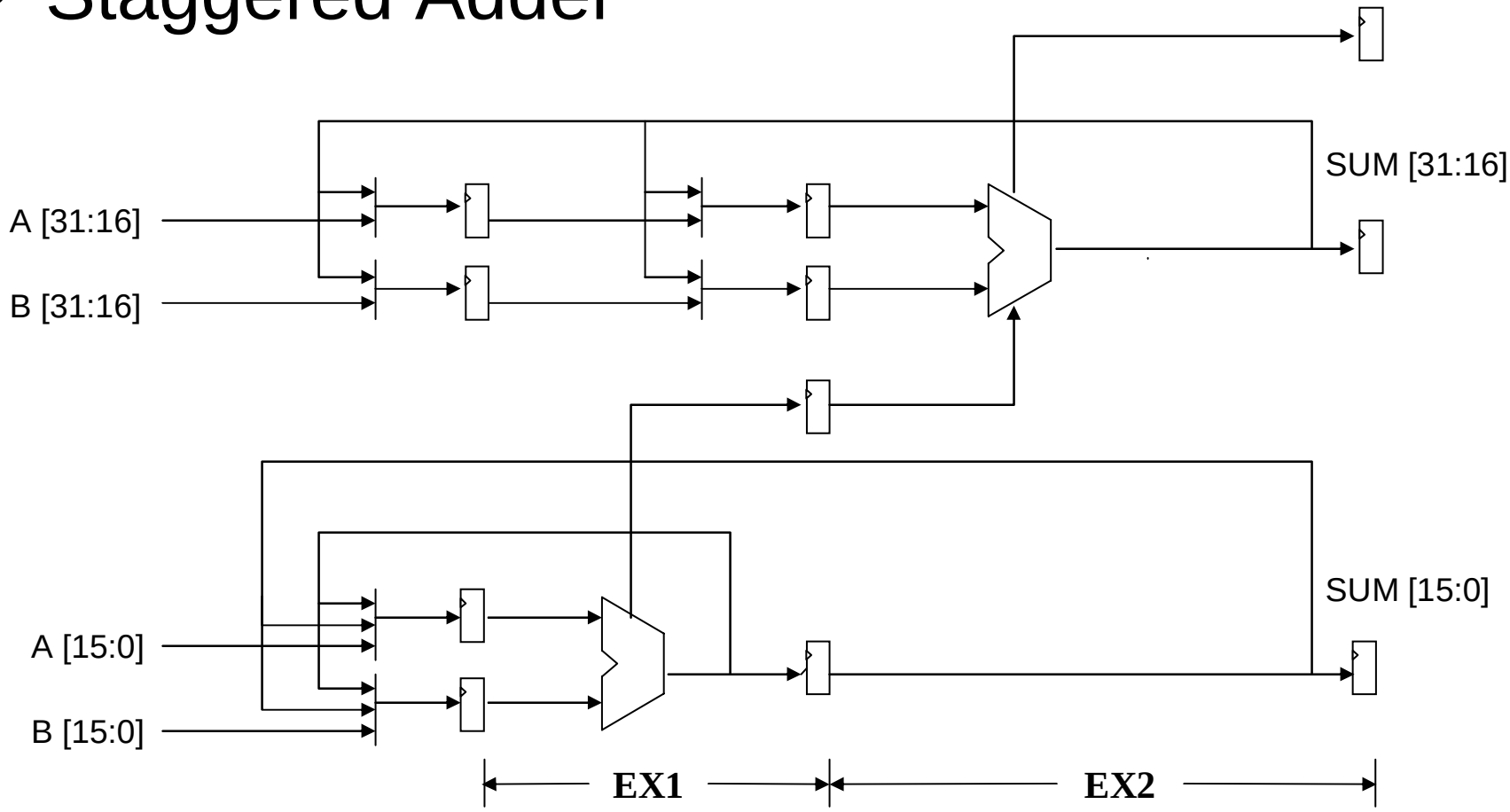
Future Work

- Design a DPS-enabled deep pipeline
- Integrate Wattch power models

  [D. Brooks, V. Tiwari, and M. Martonosi, ISCA-27]
- Investigate interaction between fetch gating and DPS

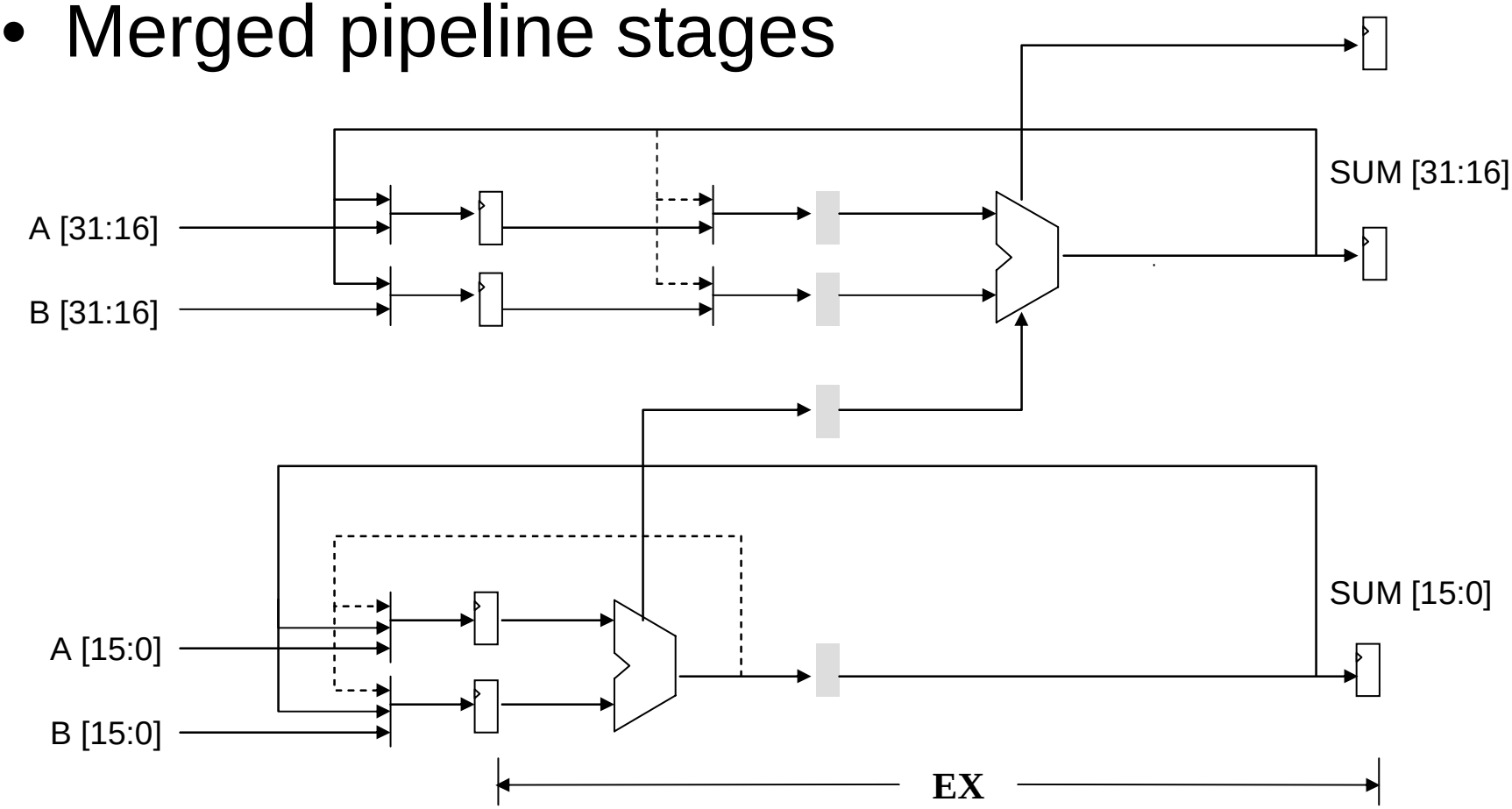
Design Example

- Staggered Adder

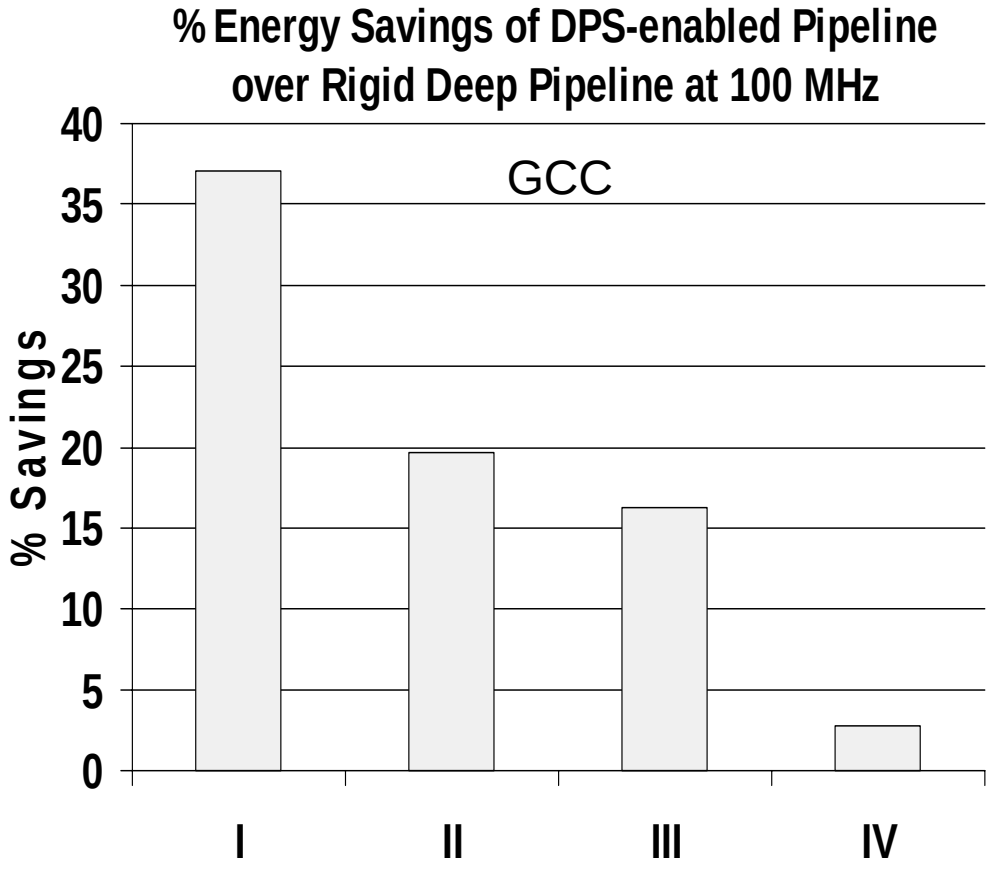


Design Example

- Merged pipeline stages



Preliminary Results



- **I: No Clock Gating, Real Branch Prediction**
- **II: No Clock Gating, Oracle Branch Prediction**
- **III: Perfect Clock Gating, Real Branch Prediction**
- **IV: Perfect Clock Gating, Oracle Branch Prediction**