Design of Energy/Quality Scalable Hardware By Runtime Voltage Scaling and Back Biasing

Daniele Jahier Pagliari

EDA Group
Politecnico di Torino
Torino, Italy

POLITECNICO
DI TORINO

leti
ceatech

TECHNOLOGY
RESEARCH
INSTITUTE

2nd IWES
September 8th , 2017, Rome, Italy

The EDA Group

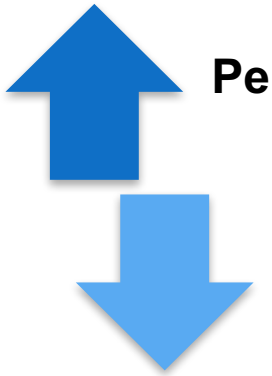
- Electronic Design Automation
  - 7 Faculty members
    - Enrico Macii, Massimo Poncino, Alberto Macii, Andrea Acquaviva
    - Elisa Ficarra, Andrea Calimera , Santa Di Cataldo, Sara Vinco
  - 4 post-doc researchers
  - ~10+ Ph.D. students & Research Assistants
- Three main areas of research:
  - *EDA (energy efficiency, EES, etc.)*
  - *Technologies for Smart Cities (Buildings, Districts, etc.)*
  - *Bioinformatics*
- Strong record of EU funded projects
  - 30+ in the last 10 years.

Outline

- Introduction
- Background and Motivation
- Dynamic V_{DD}/V_{BB}/Accuracy Tuning
- Experimental Results
- Conclusions and Future Work

Introduction

- IoT devices trends:

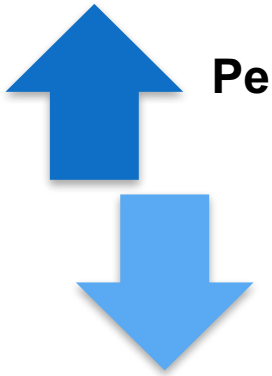


Performance Demands

Energy Budget

Introduction

- IoT devices trends:



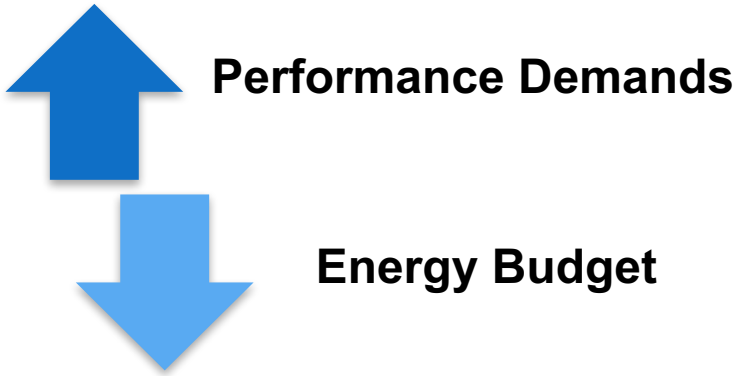
Performance Demands

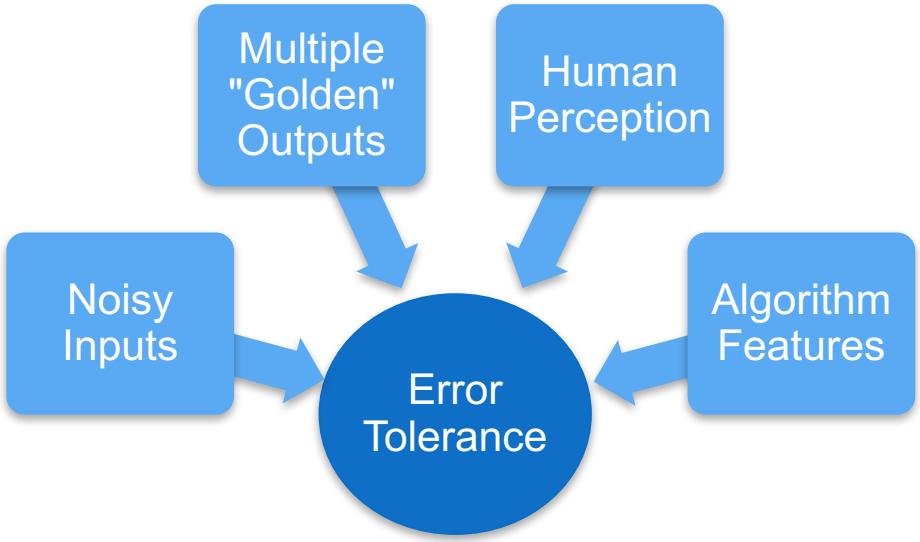
Energy Budget

- Many emerging applications are **error tolerant** (or *error resilient*):
  - Recognition, Mining and Synthesis (RMS) domains

Introduction

- IoT devices trends:

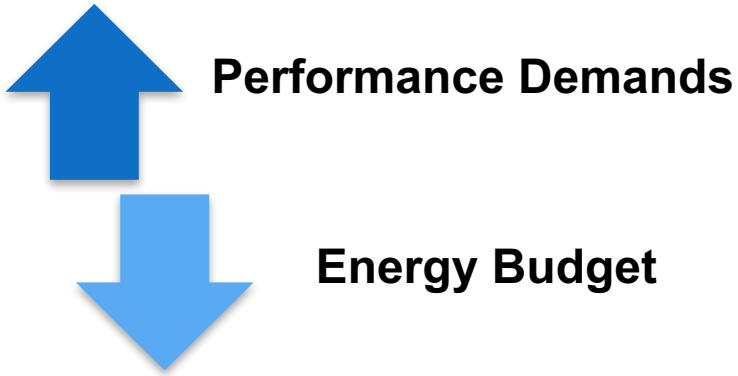


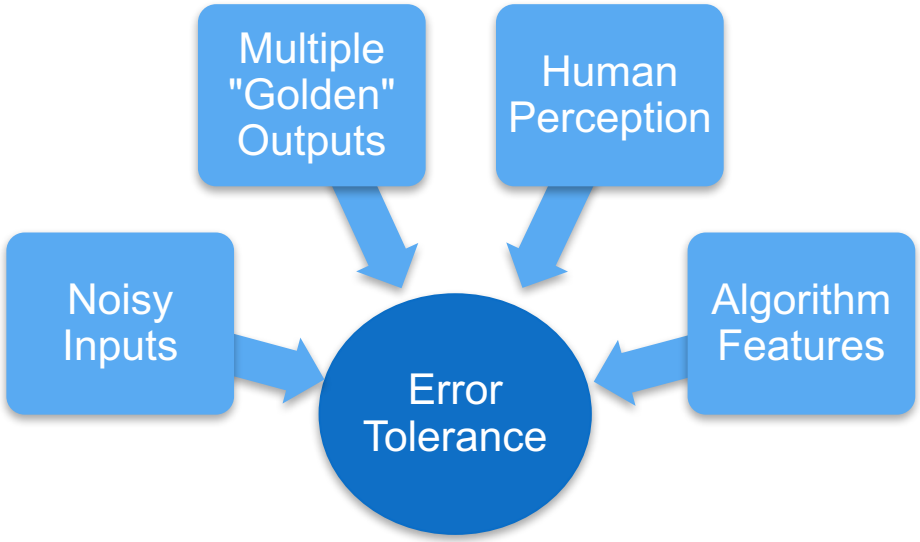


- Many emerging applications are **error tolerant** (or *error resilient*):
  - Recognition, Mining and Synthesis (RMS) domains

Introduction

- IoT devices trends:



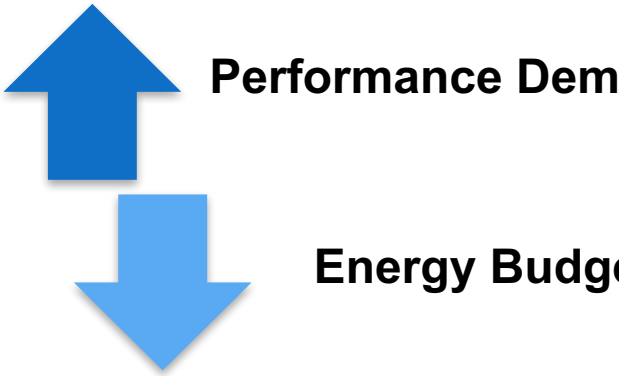


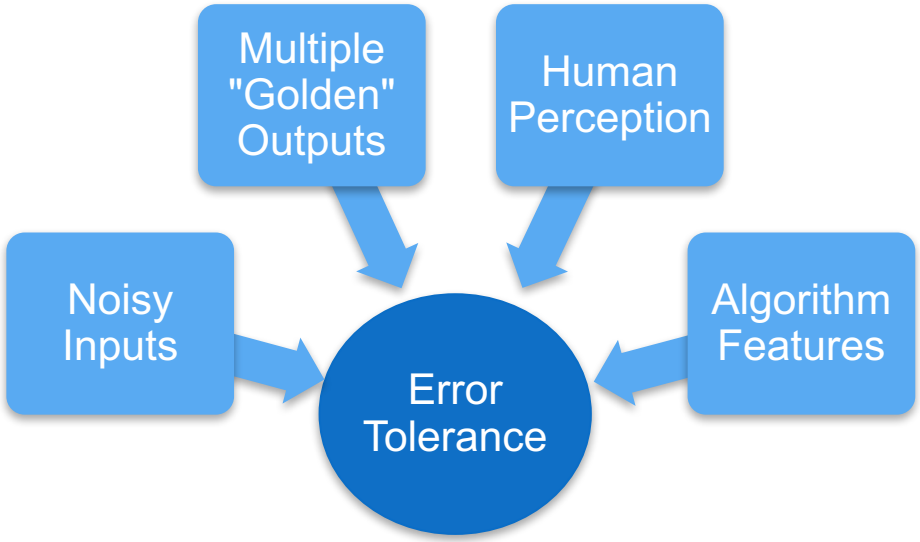
- Many emerging applications are **error tolerant** (or *error resilient*):
  - Recognition, Mining and Synthesis (RMS) domains

Approximate Paradigm: Tradeoff ***energy consumption*** and ***output quality*** leveraging applications error tolerance.

Introduction

- IoT devices trends:





- Many emerging applications are **error tolerant** (or *error resilient*):
  - Recognition, Mining and Synthesis (RMS) domains

Approximate Paradigm: Tradeoff ***energy consumption*** and ***output quality*** leveraging applications error tolerance.

- **Two main approaches:**
  - **Design-time Approximations**
  - ***Quality-Configurable Systems (QCS)***

Background - Functional Units

1. **Approximate circuits:**
   - Mostly adders and multipliers

   Kyaw, Goh and Yeo, EDSSC'10, Huang, Lach and Robins, DAC'12, Farshchi, Saeed and Fakhraie, CADS'13, Jiang, Han and Lombardi, GLSVLSI'15, Bhardwaj, Mane and Henkel, ISQED'15, etc.

2. **Approximate synthesis:**
   - Generalization of the previous techniques to any netlist

   Shin and Gupta, ATS'08, Venkataramani et al, DAC'12, Miao, Gerstlauer and Orshansky, ICCAD'13, Jahier Pagliari et al, ICCD'15,etc.

3. **Quality-configurable circuit architectures:**
   - Arithmetic units

   De la Guya Solaz, Han, Conway, IEEE TCAS'11, Kahng and Kang, DAC'12, Ye et al, ICCAD'13, Liu, Han and Lombardi, DATE'14, etc.
   - Voltage scalable meta-functions

   Mohapatra, Chippa, Raghunathan and Roy, DATE'11

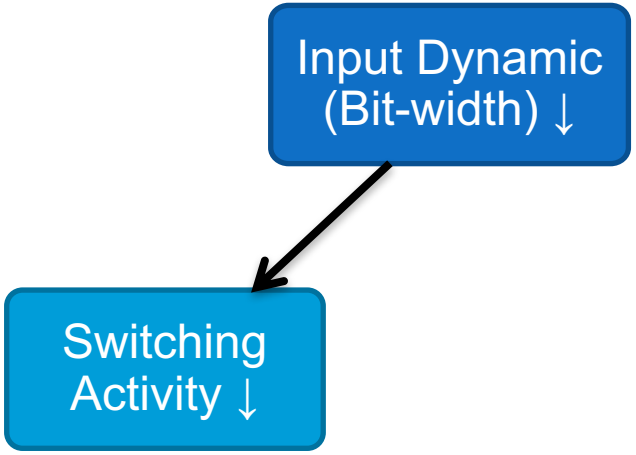
4. **Dynamic Voltage and Accuracy Scaling (DVAS):**
   - Use technological knobs only (no design modifications)

   Moons and Verhelst, ISLPED'15, Moons et al, ISSCC'17, etc.

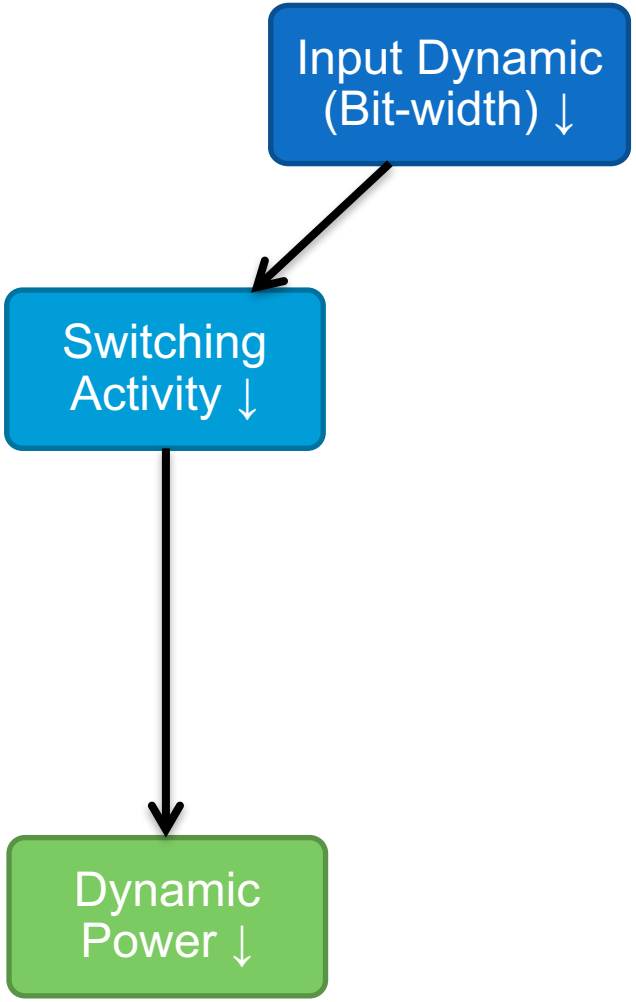
Background - DVAS

Input Dynamic
(Bit-width) ↓

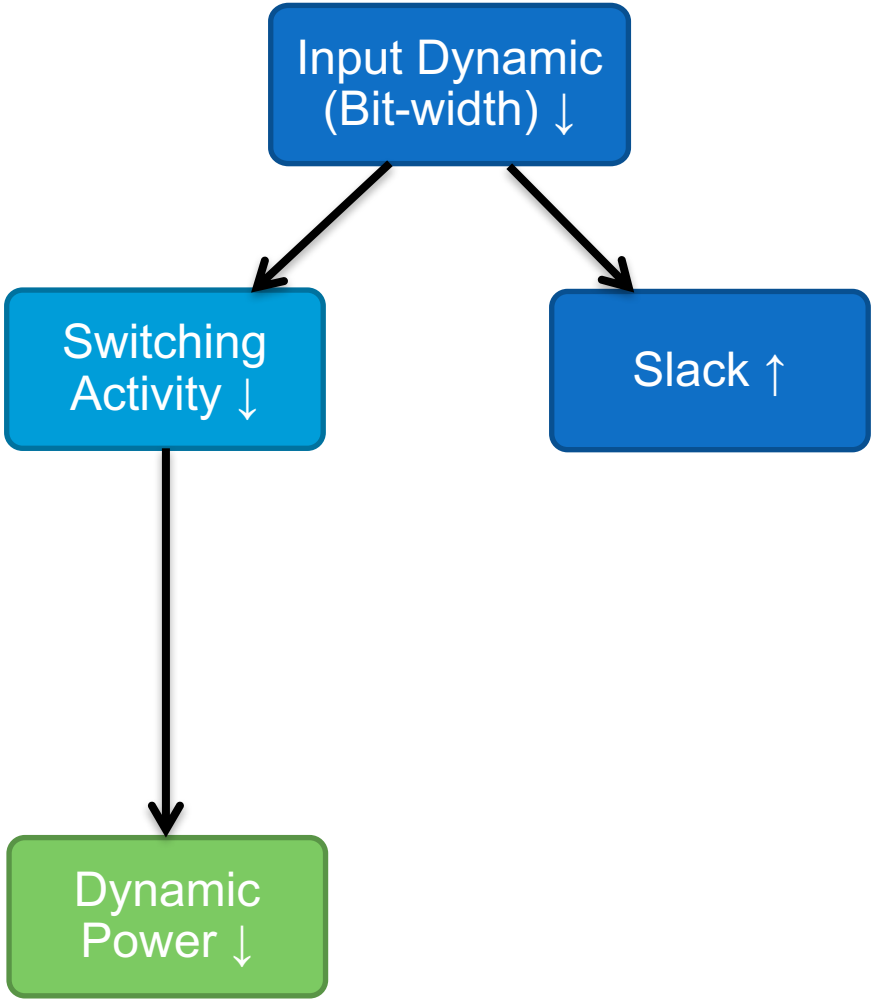
Background - DVAS



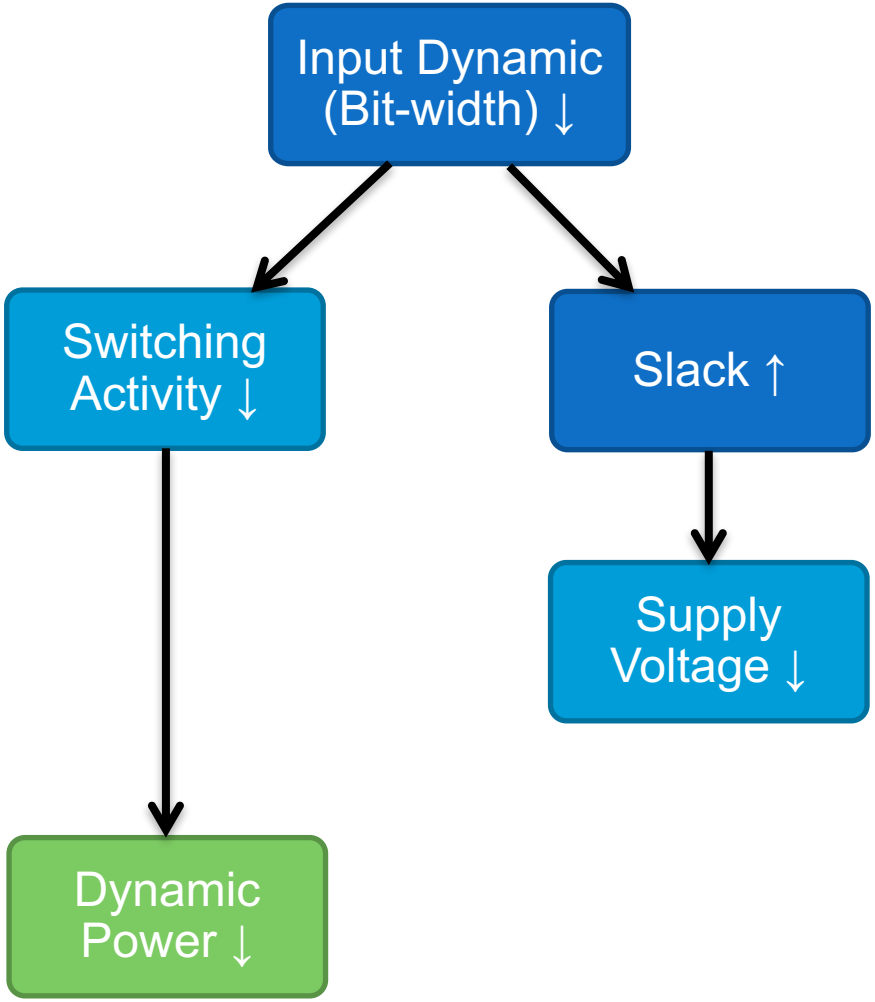
Background - DVAS



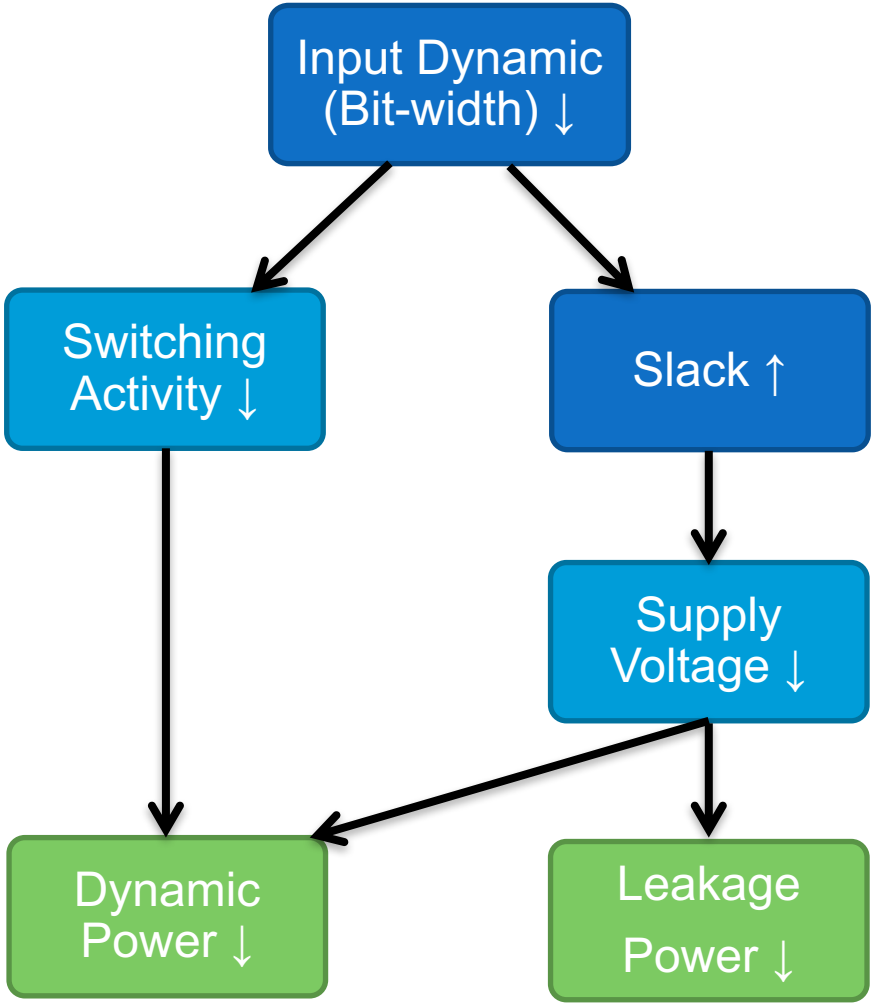
Background - DVAS



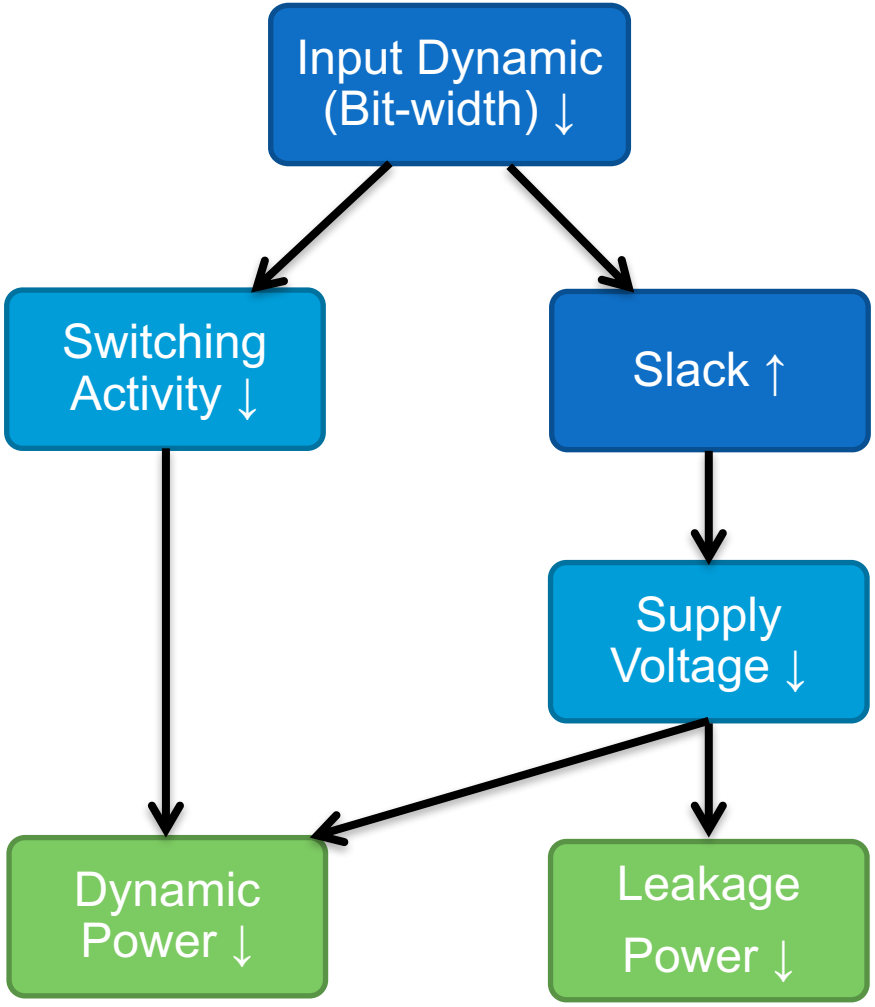
Background - DVAS

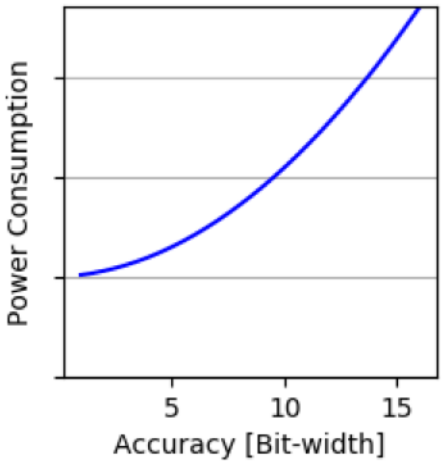


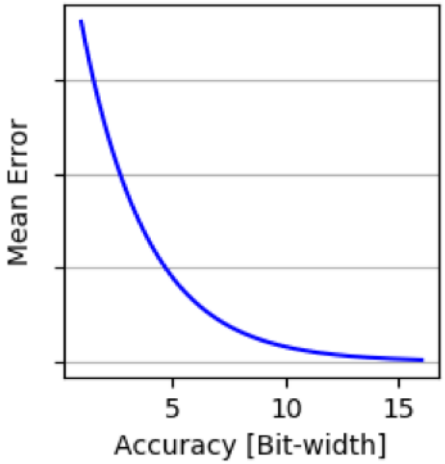
Background - DVAS



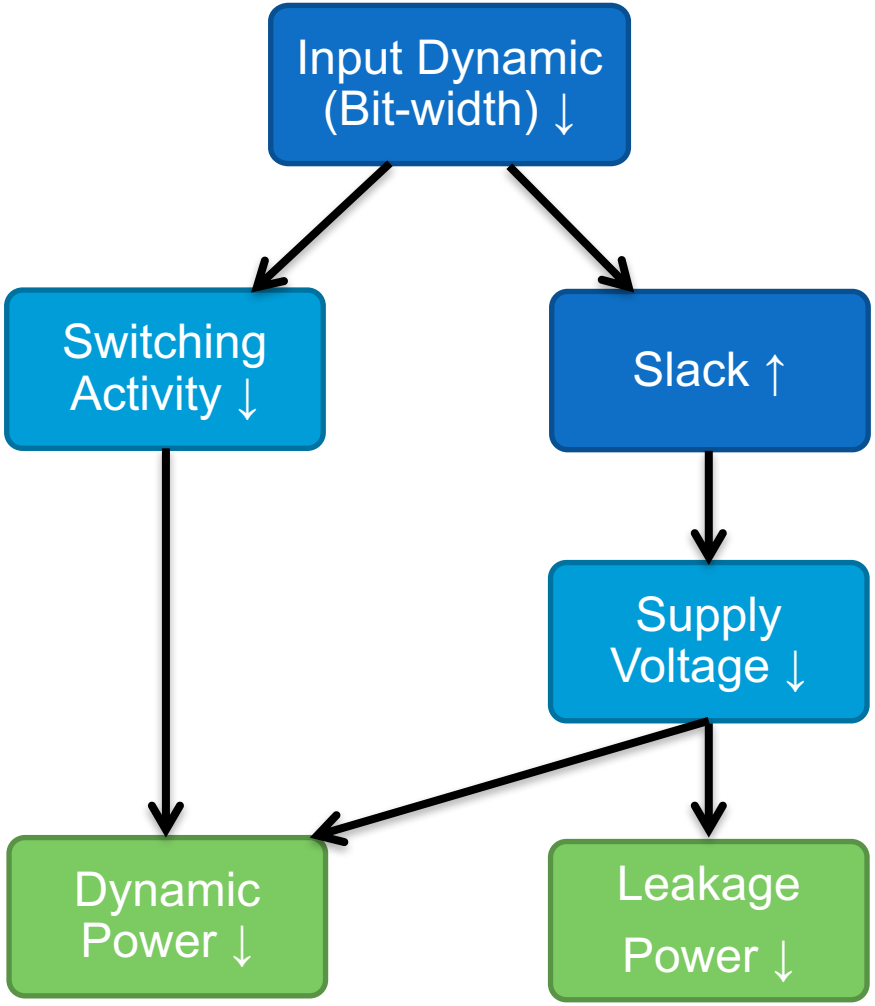
Background - DVAS

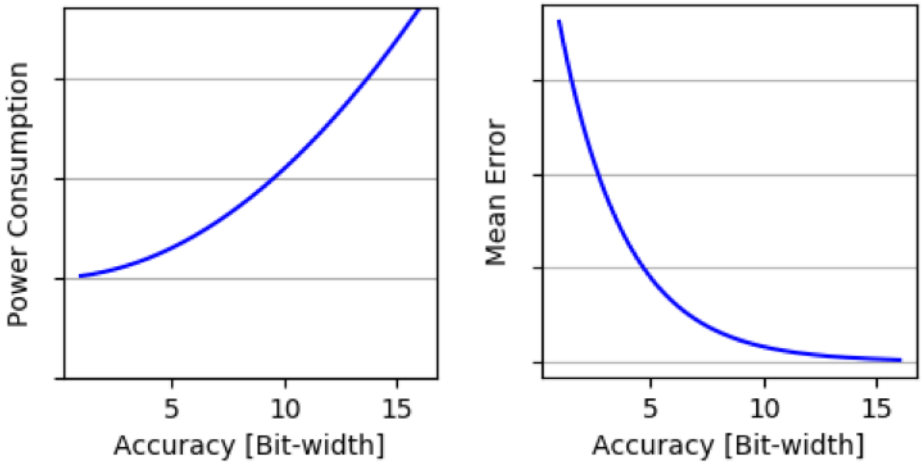






Background - DVAS





- **Pros of DVAS:**
  - No overheads at max. accuracy (*)
  - Bounded error
  - Many quality configurations (N-bit)
  - Architecture-independent

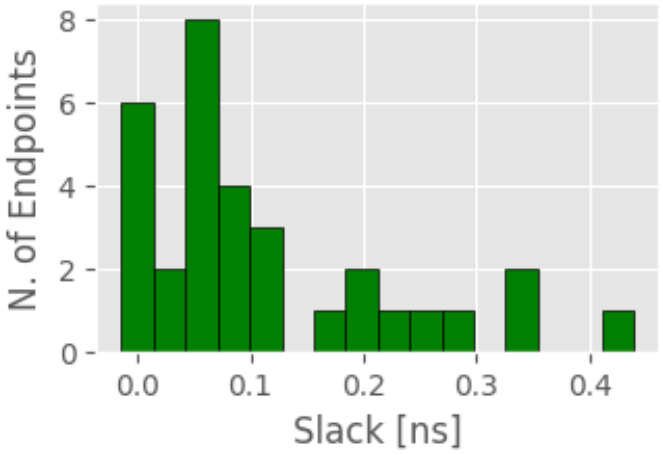
Motivation

- **Main Limitation of DVAS:**
  "Wall-of-Slack" phenomenon:
  - Synthesis optimizes long paths for **timing**, short ones for **area and power**
  - Most paths become "almost-critical"

Motivation

- **Main Limitation of DVAS:**
  "Wall-of-Slack" phenomenon:
  - Synthesis optimizes long paths for **timing**, short ones for **area and power**
  - Most paths become "almost-critical"

- **Example:**
  - Booth multiplier endpoint histogram.

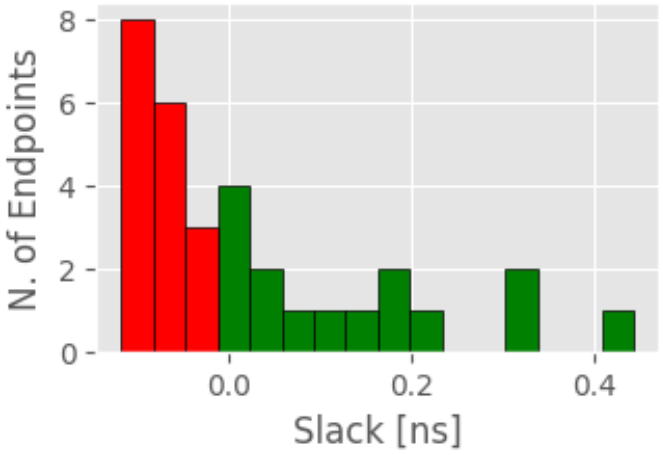


1.00V

Motivation

- **Main Limitation of DVAS:**
  "Wall-of-Slack" phenomenon:
  - Synthesis optimizes long paths for **timing**, short ones for **area and power**
  - Most paths become "almost-critical"
  - When V_{DD} is scaled the number of usable bits decreases rapidly

- **Example:**
  - Booth multiplier endpoint histogram.

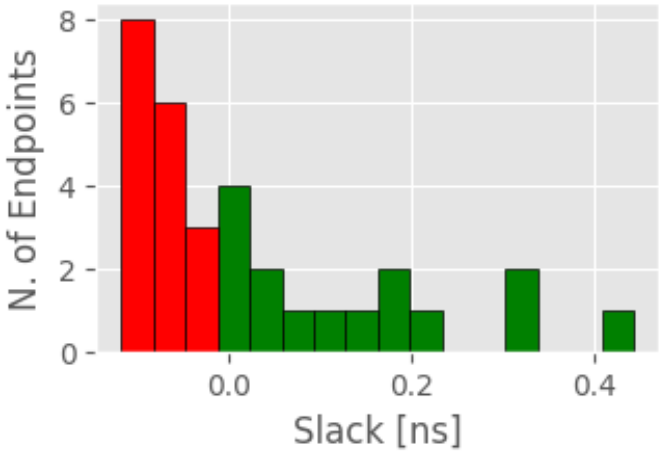


0.90V

Motivation

- **Main Limitation of DVAS:**
  "Wall-of-Slack" phenomenon:
  - Synthesis optimizes long paths for **timing**, short ones for **area and power**
  - Most paths become "almost-critical"
  - When V_{DD} is scaled the number of usable bits decreases rapidly

- **Example:**
  - Booth multiplier endpoint histogram.



0.90V

Useful bit-width configurations require $V_{DD} \cong V_{DD,NOM}$

Motivation (cont'd)

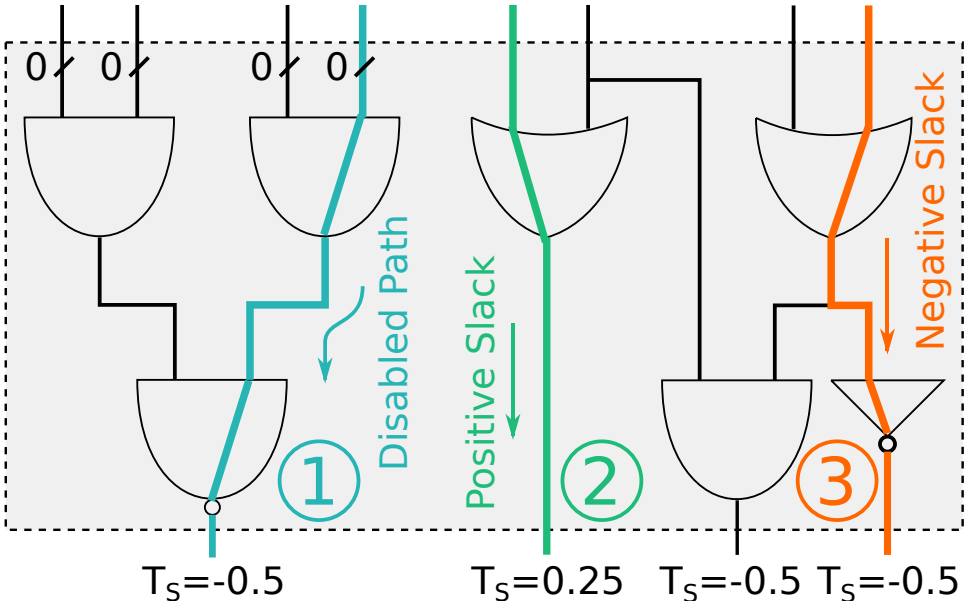
Contrasting the "Wall of Slack":

- **Solution 1:** modify synthesis constraints.
  - Overhead in area and **power at maximum accuracy**.

Motivation (cont'd)

Contrasting the "Wall of Slack":

- **Solution 1:** modify synthesis constraints.
  - Overhead in area and **power at maximum accuracy**.
- **Solution 2:** **finer-grain power/delay tuning**
  - **Key:** in reduced accuracy "modes", not all **paths** of the circuit require the same "speed"



Fine-grain power/delay tuning

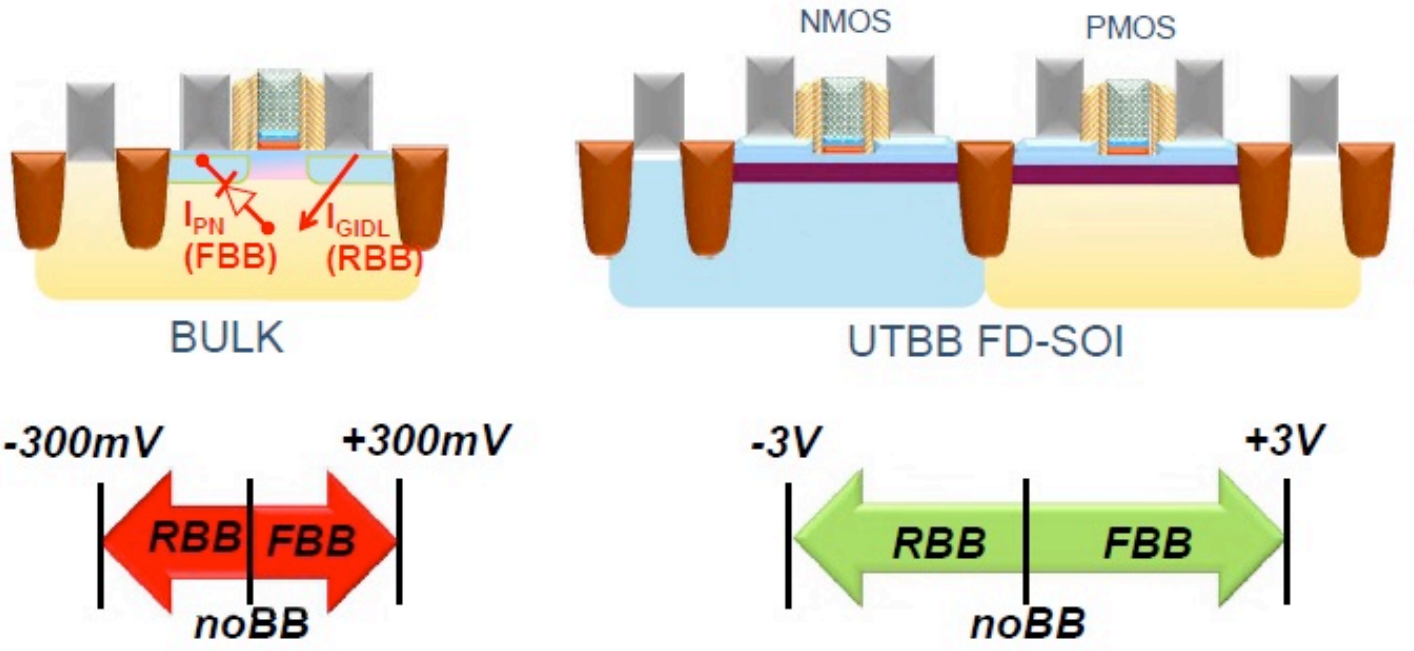
- **Possible solution: multiple VDD**
  - Requires level shifters
  - Excessive power overheads for a single FU

Fine-grain power/delay tuning

- **Possible solution: multiple VDD**
  - Requires level shifters
  - Excessive power overheads for a single FU
- **Our solution: combine DVAS** with **FDSOI's Back Bias**

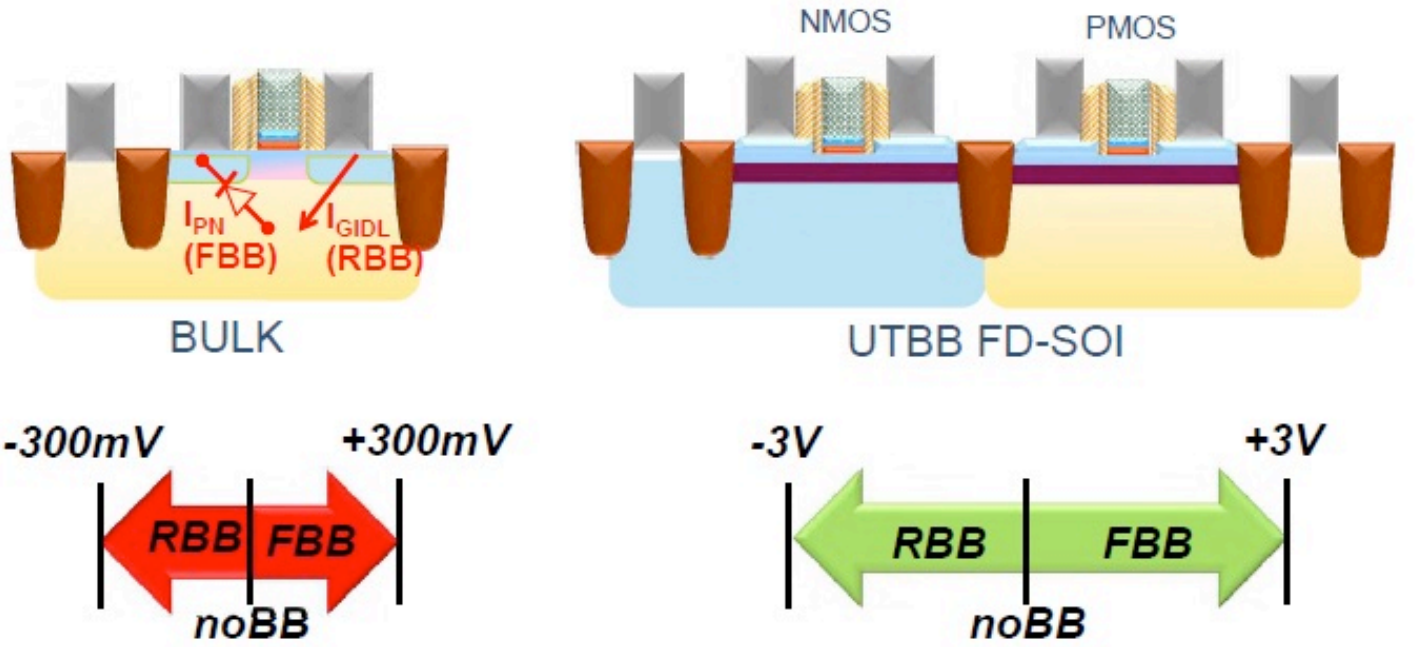
Fine-grain power/delay tuning

- **Possible solution: multiple VDD**
  - Requires level shifters
  - Excessive power overheads for a single FU
- **Our solution: combine DVAS with FDSOI's Back Bias**



Fine-grain power/delay tuning

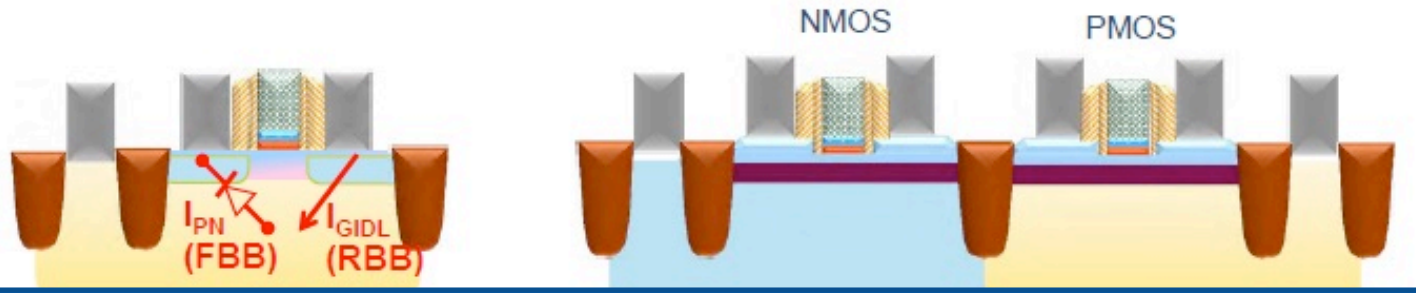
- **Possible solution: multiple VDD**
  - Requires level shifters
  - Excessive power overheads for a single FU
- **Our solution: combine DVAS with FDSOI's Back Bias**



- **Fine-grain threshold voltage (V_{th}) tuning in addition to V_{DD} assignment**

Fine-grain power/delay tuning

- **Possible solution: multiple VDD**
  - Requires level shifters
  - Excessive power overheads for a single FU
- **Our solution: combine DVAS with FDSOI's Back Bias**



Advantages:
- Fine-grain speed/power control
- V_{DD} possibly shared with other FUs
- No level shifters; *Well insulation trenches* (area overhead only)

- **assignment**

Dynamic VDD/VBB/Accuracy Tuning

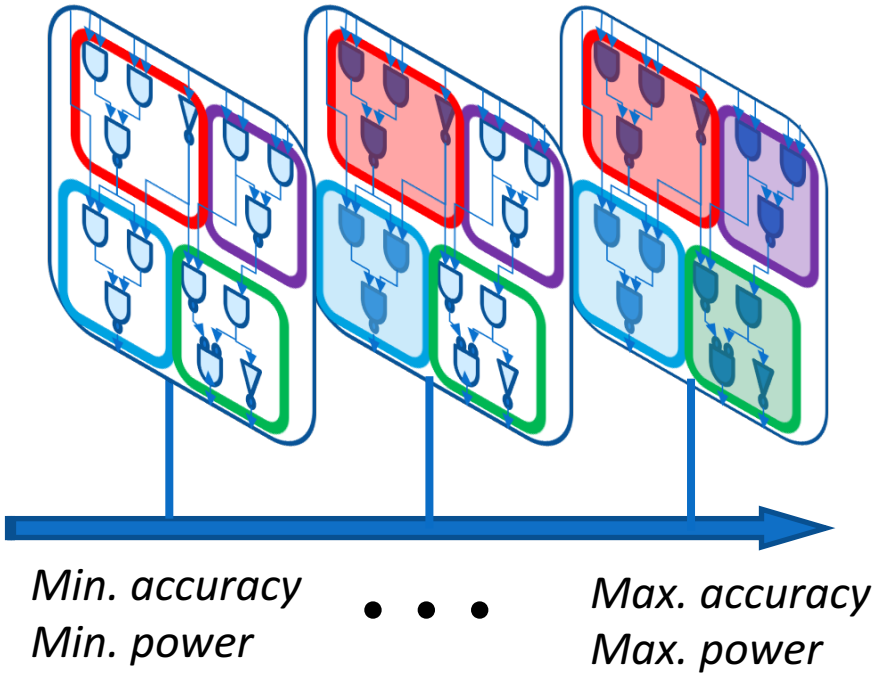
Issue with V_{BB} assignment:

- Cannot apply independent V_{BB} to each cell
- Partition in **V_{BB} domains** is required

Dynamic VDD/VBB/Accuracy Tuning

Issue with V_{BB} assignment:

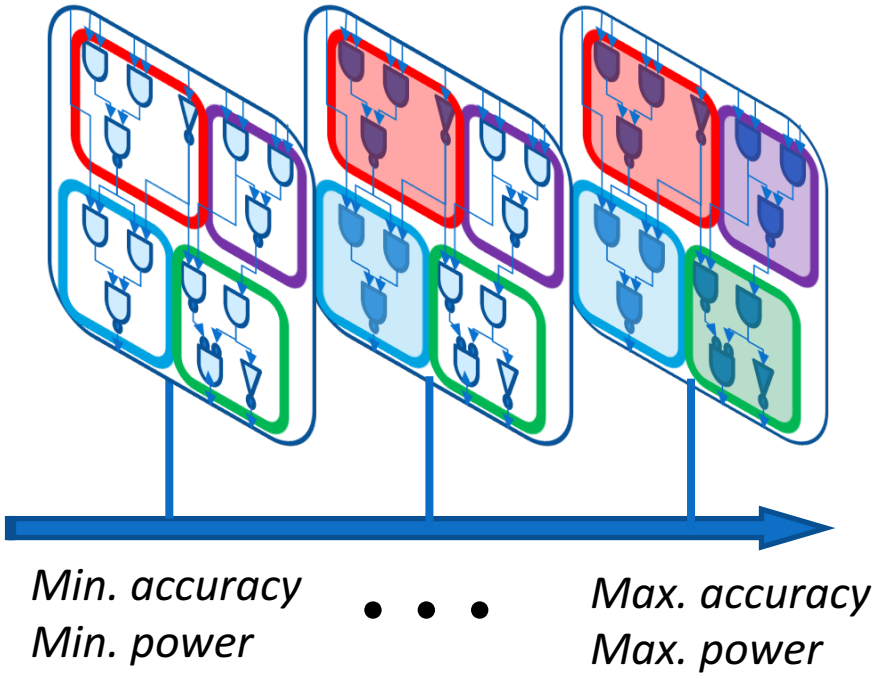
- Cannot apply independent V_{BB} to each cell
- Partition in **V_{BB} domains** is required



Dynamic VDD/VBB/Accuracy Tuning

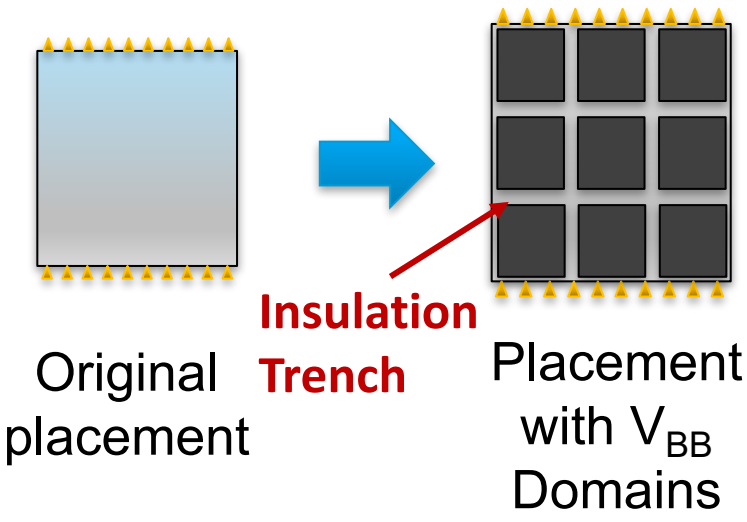
Issue with V_{BB} assignment:

- Cannot apply independent V_{BB} to each cell
- Partition in **V_{BB} domains** is required



Proposed partitioning:

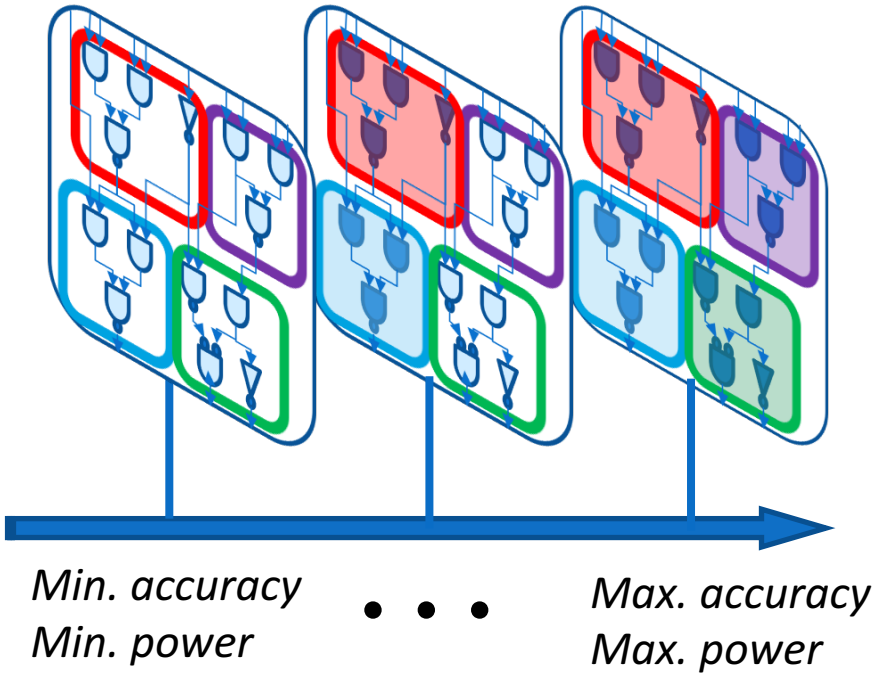
Regular Tiling



Dynamic VDD/VBB/Accuracy Tuning

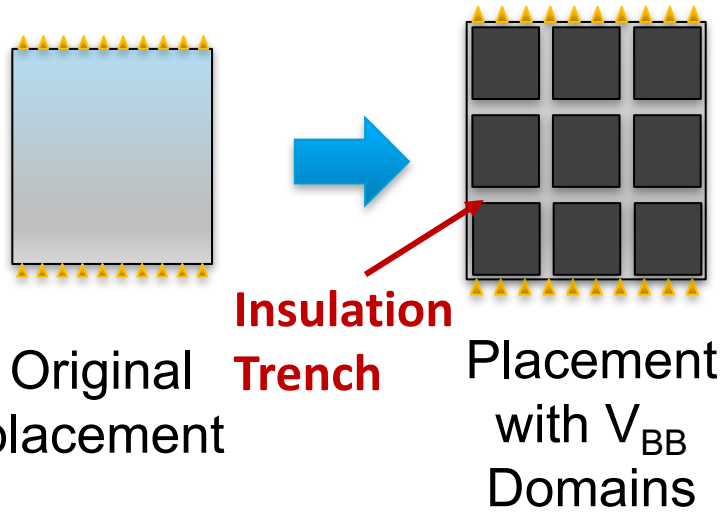
Issue with V_{BB} assignment:

- Cannot apply independent V_{BB} to each cell
- Partition in **V_{BB} domains** is required



Proposed partitioning:

Regular Tiling



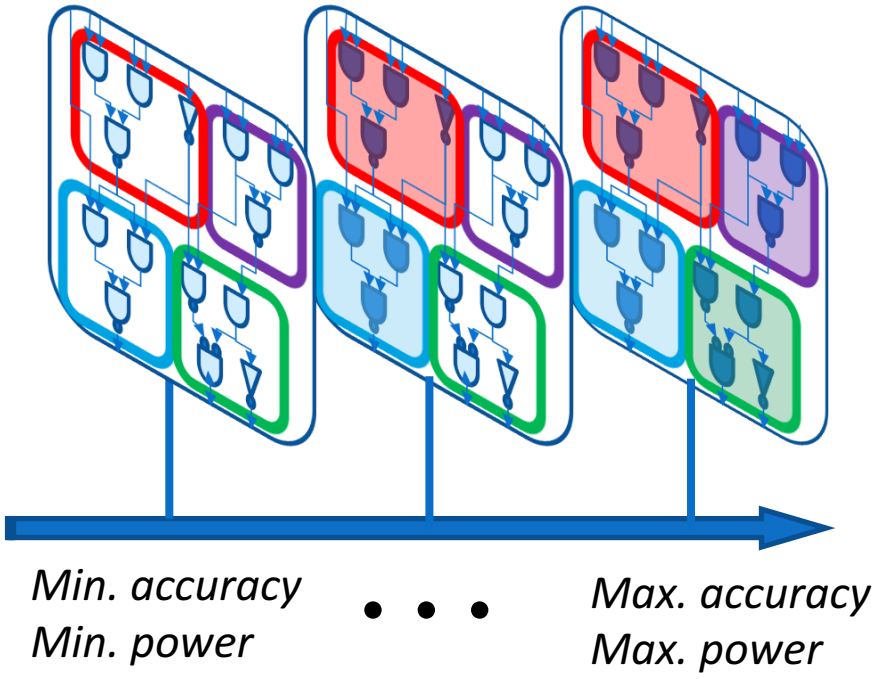
Pros:

- Regularity of design
- Easy to incorporate in EDA flow
- Minimal displacement of cells

Dynamic VDD/VBB/Accuracy Tuning

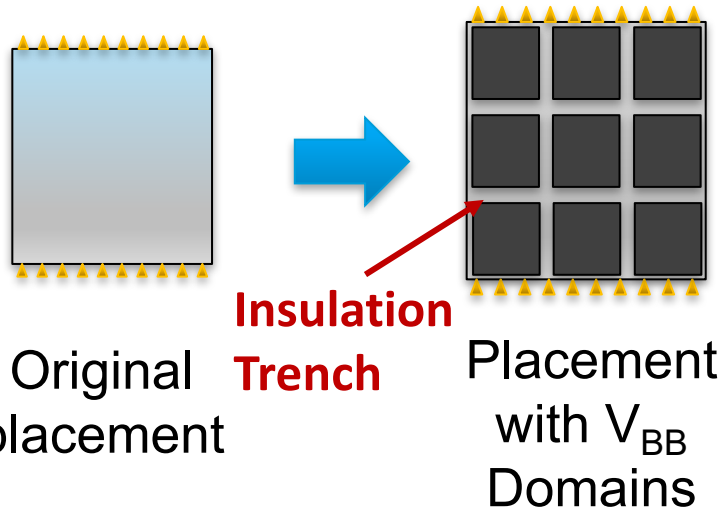
Issue with V_{BB} assignment:

- Cannot apply independent V_{BB} to each cell
- Partition in **V_{BB} domains** is required



Proposed partitioning:

Regular Tiling



Pros:

- Regularity of design
- Easy to incorporate in EDA flow
- Minimal displacement of cells

Minimal timing, area and power overheads at maximum accuracy.

Experimental Results

Designs:

- Booth **multiplier**
- **FFT** Butterfly unit
- 30-tap **FIR** filter
- 16-bit fixed-point implementations

Experimental Results

Designs:

- Booth **multiplier**
- **FFT** Butterfly unit
- 30-tap **FIR** filter
- 16-bit fixed-point implementations

Operating Conditions:

- V_{DD} = [0.6V, 0.7V,…1.0V]
- Forward BB: V_{BB} = ±1.1V (N-Well/P-Well)

Experimental Results

Designs:

- Booth **multiplier**
- **FFT** Butterfly unit
- 30-tap **FIR** filter
- 16-bit fixed-point implementations

Operating Conditions:

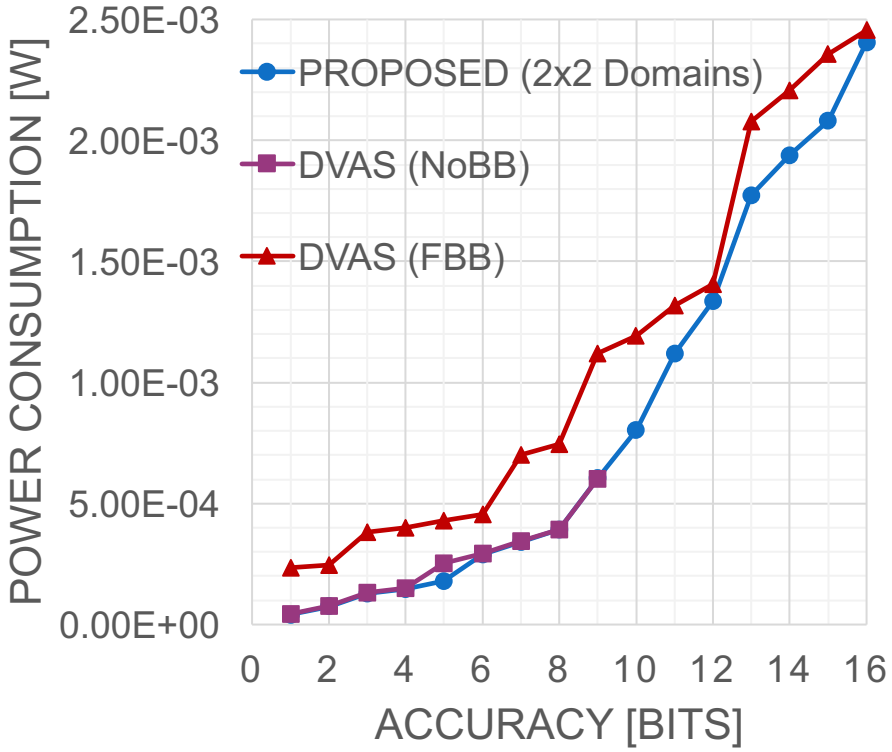
- V_{DD} = [0.6V, 0.7V,…1.0V]
- Forward BB: V_{BB} = ±1.1V (N-Well/P-Well)

Design	Area [mm^2]	Clock Freq. [GHz]	# of V_{BB} Domains
Booth	2.59e-03	1.25	2 x 2
Butterfly	7.71e-03	1.00	3 x 3
FIR	9.10e-03	0.75	3 x 3

Comparison with DVAS

Booth Multiplier

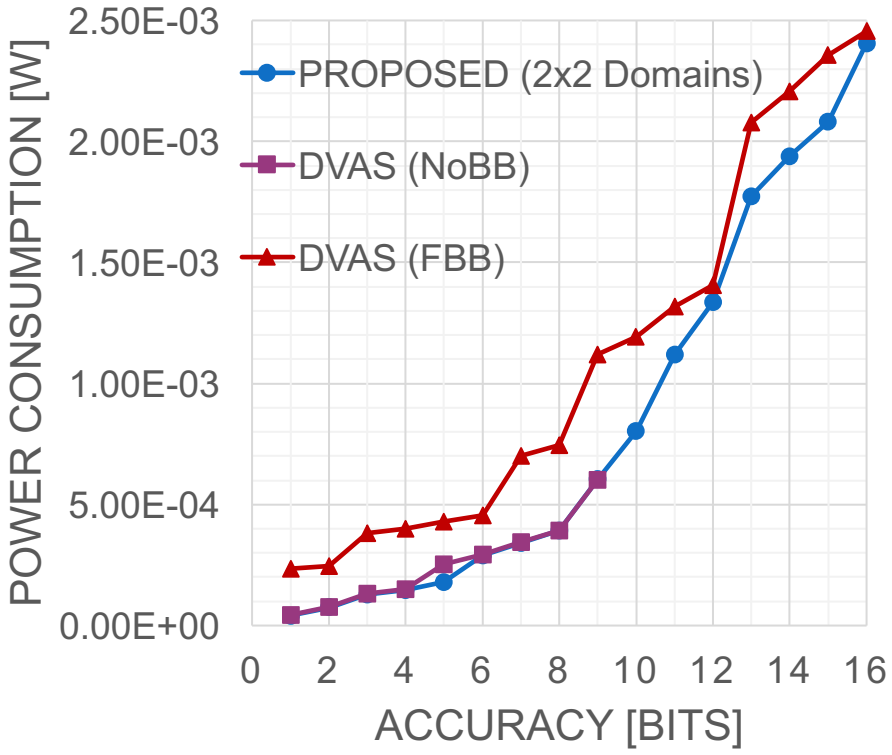
- **Plots:** Minimum power configuration for each accuracy
- Combining (global) V_{DD} scaling and fine-grain back-biasing



Comparison with DVAS

Booth Multiplier

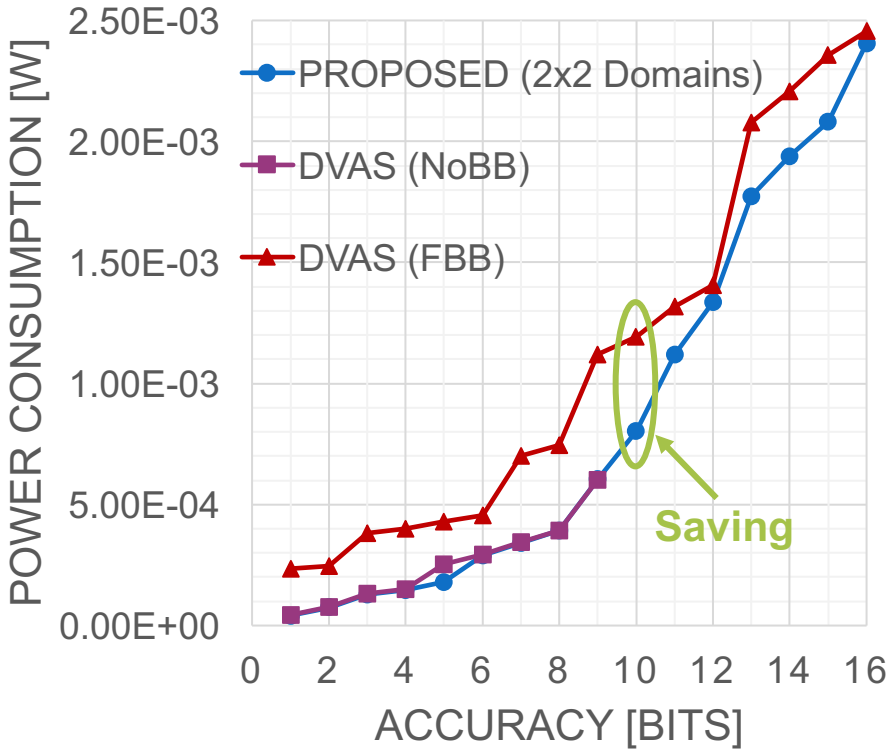
- **Plots:** Minimum power configuration for each accuracy
- Combining (global) V_{DD} scaling and fine-grain back-biasing
- **Comparison:**
  - DVAS with No Back Biasing (**NoBB**)
  - DVAS with **FBB** in the entire circuit



Comparison with DVAS

Booth Multiplier

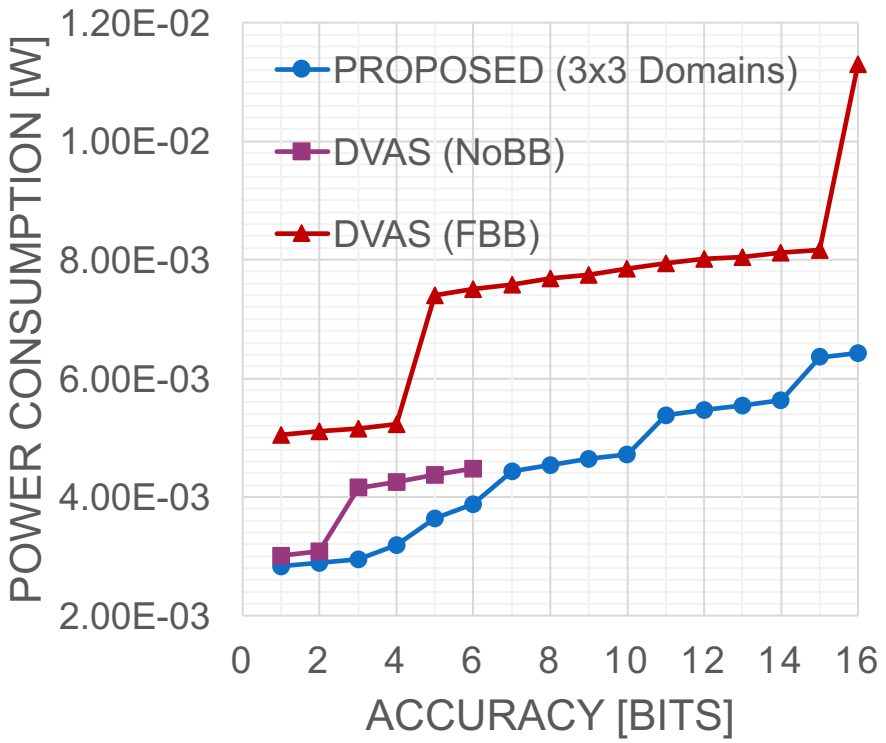
- **Plots:** Minimum power configuration for each accuracy
- Combining (global) V_{DD} scaling and fine-grain back-biasing
- **Comparison:**
  - DVAS with No Back Biasing (**NoBB**)
  - DVAS with **FBB** in the entire circuit
- **32.7% Saving w.r.t. DVAS @ 10-bit accuracy!**



Comparison with DVAS

FIR Filter

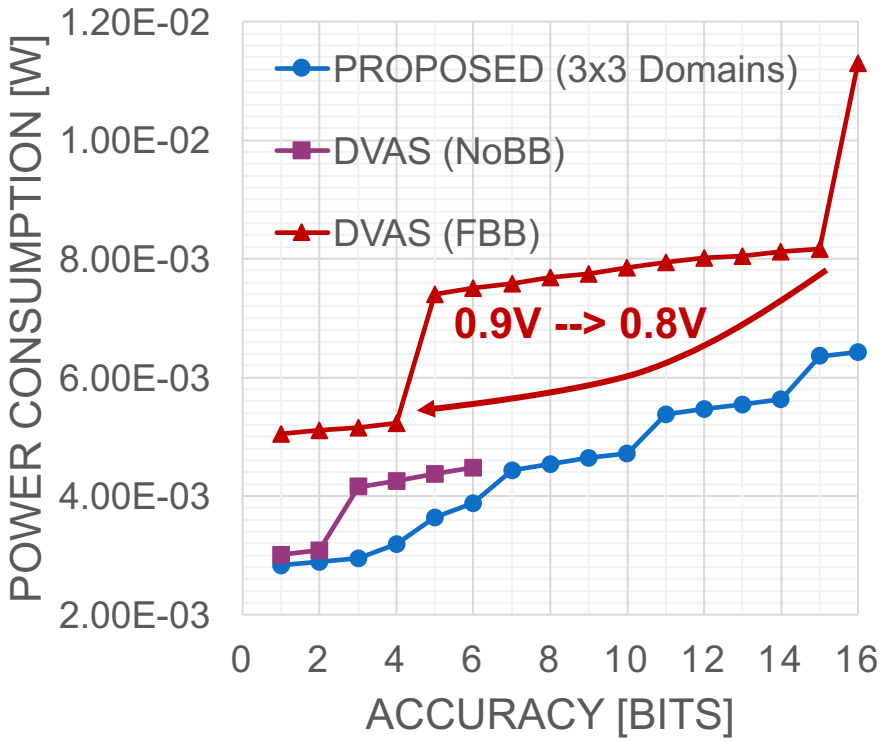
- "Wall-of-Slack" clearly visible
- Maximum DVAS + FBB accuracy (without violations):
  - 15-bit @ 0.9V
  - **Only 4-bit @ 0.8V!**



Comparison with DVAS

FIR Filter

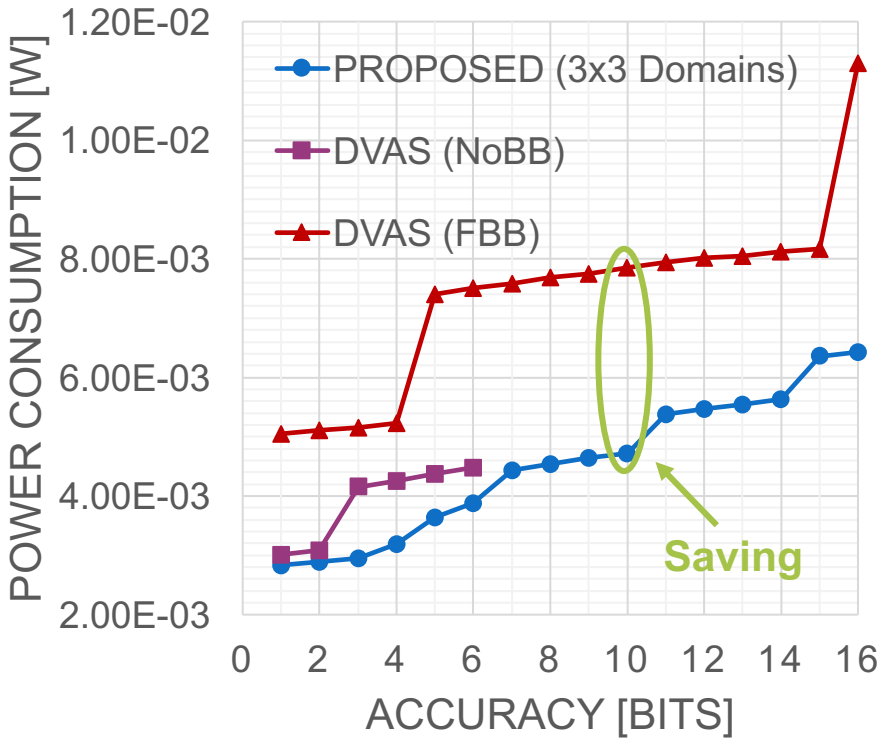
- "Wall-of-Slack" clearly visible
- Maximum DVAS + FBB accuracy (without violations):
  - 15-bit @ 0.9V
  - **Only 4-bit @ 0.8V!**



Comparison with DVAS

FIR Filter

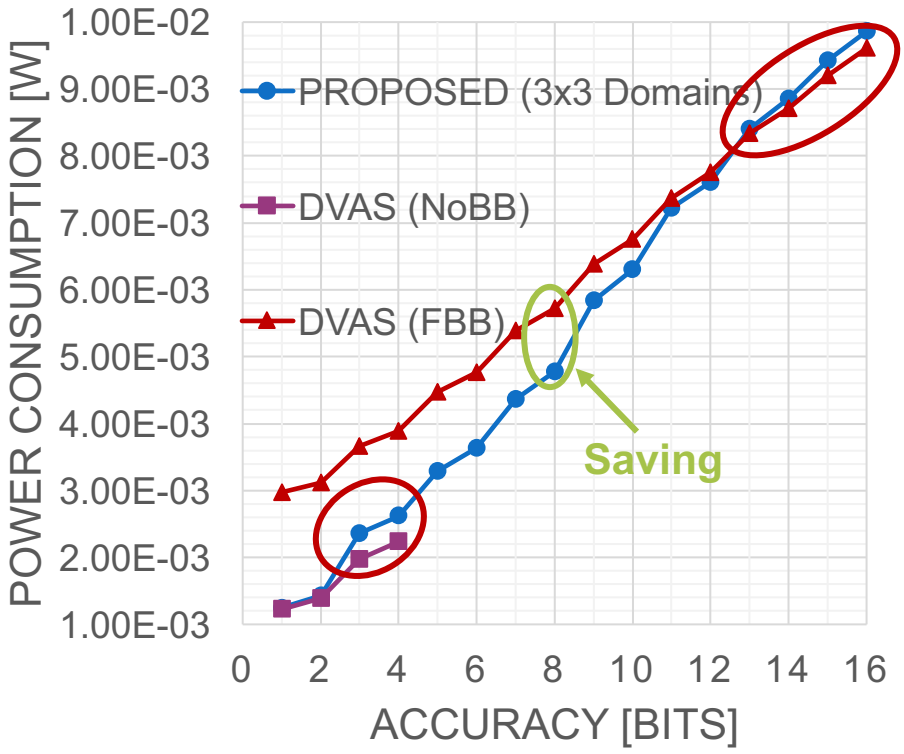
- "Wall-of-Slack" clearly visible
- Maximum DVAS + FBB accuracy (without violations):
  - 15-bit @ 0.9V
  - **Only 4-bit @ 0.8V!**
- **39.9% Saving w.r.t. DVAS @ 10-bit accuracy!**



Comparison with DVAS

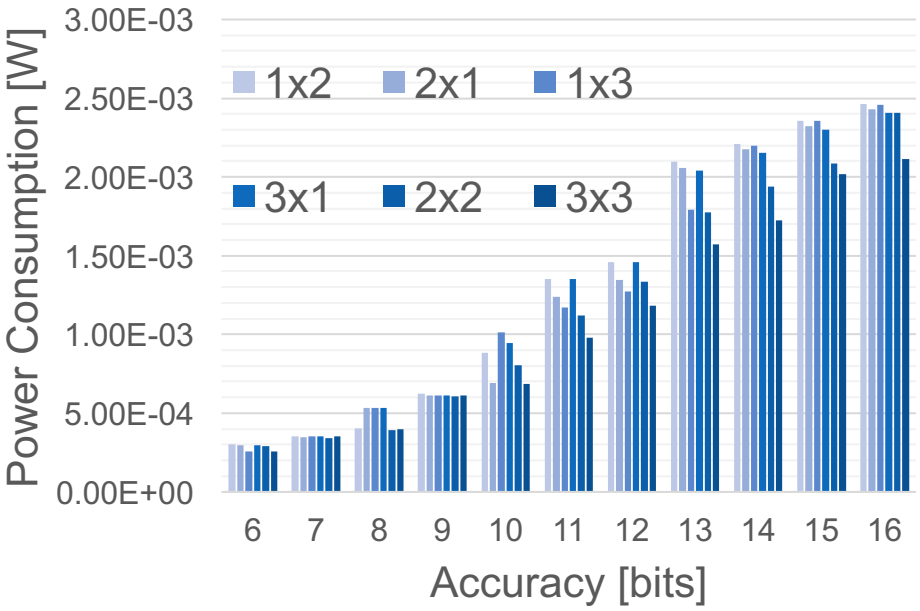
FFT Butterfly

- Large number of V_{BB} domains (3 x 3) compared to relatively small circuit area
- Power **overheads** more significant
- Also, “Wall-of-Slack” less visible (circuit probably under constrained)
- **Still 16.5% saving w.r.t. DVAS @ 8-bit!**



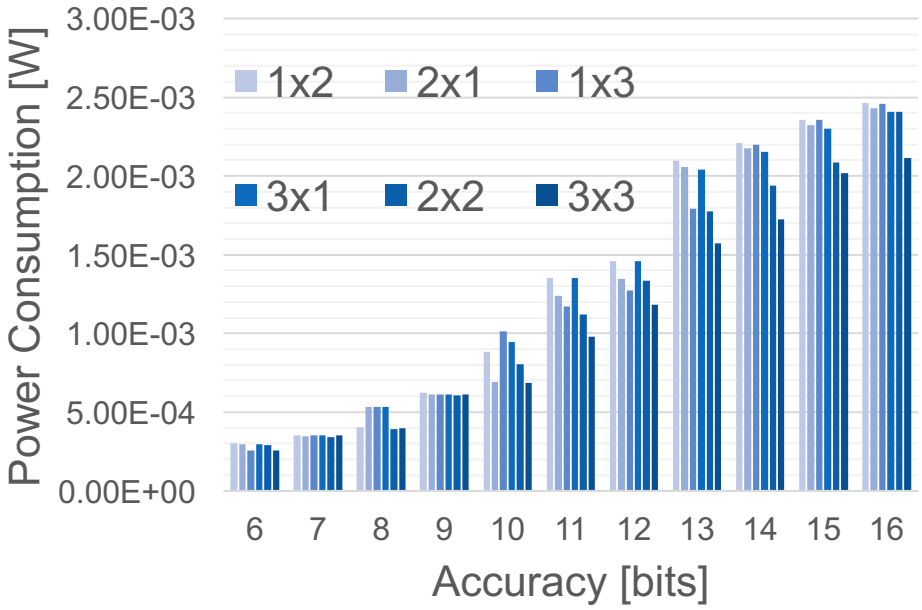
Impact of V_{BB} Domains

- **Number of V_{BB} domains vs power saving (Booth Mul.):**

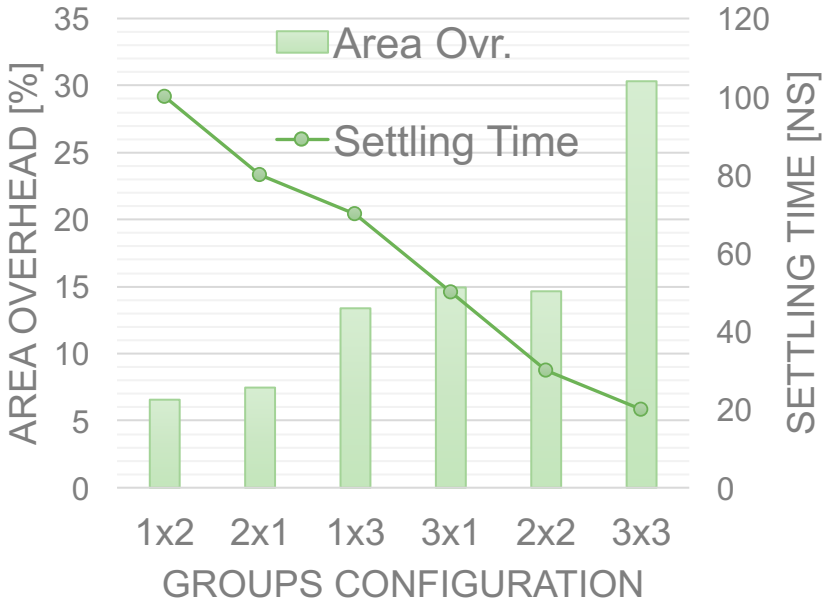


Impact of V_{BB} Domains

- **Number of V_{BB} domains vs power saving (Booth Mul.):**



- **Number of V_{BB} domains vs overheads (Booth Mul.):**



Conclusions and Future Work

Conclusions:

- Back-Bias is an effective knob for **fine-grain delay/power tuning** in quality-configurable functional units.

Conclusions and Future Work

Conclusions:

- Back-Bias is an effective knob for **fine-grain delay/power tuning** in quality-configurable functional units.
- Combined with global V_{DD} scaling, this method **overcomes the limitations of DVAS**, by contrasting the "Wall-of-slack" phenomenon.

Conclusions and Future Work

Conclusions:

- Back-Bias is an effective knob for **fine-grain delay/power tuning** in quality-configurable functional units.
- Combined with global V_{DD} scaling, this method **overcomes the limitations of DVAS**, by contrasting the "Wall-of-slack" phenomenon.
- First ever **application of Back-Biasing to Quality Configurable Systems** (to our knowledge).

Conclusions and Future Work

Conclusions:

- Back-Bias is an effective knob for **fine-grain delay/power tuning** in quality-configurable functional units.
- Combined with global V_{DD} scaling, this method **overcomes the limitations of DVAS**, by contrasting the "Wall-of-slack" phenomenon.
- First ever **application of Back-Biasing to Quality Configurable Systems** (to our knowledge).

Future Developments:

- Devise method for **runtime update** of V_{BB} domains configurations depending on operating conditions (PVT, aging, etc.)

Conclusions and Future Work

Conclusions:

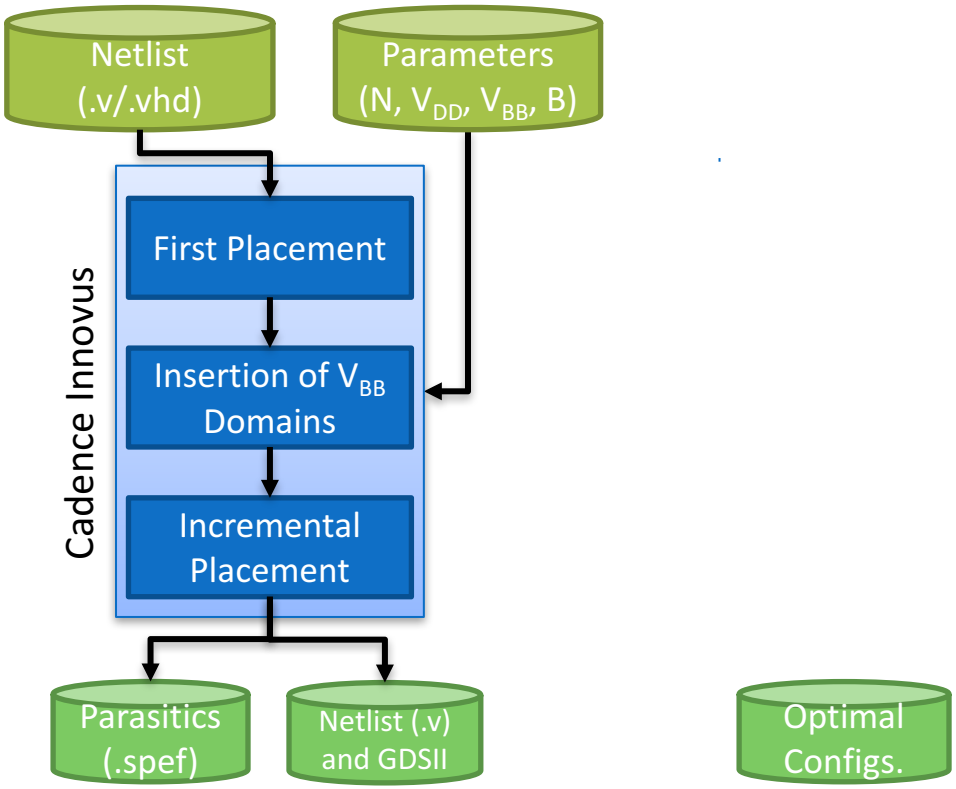
- Back-Bias is an effective knob for **fine-grain delay/power tuning** in quality-configurable functional units.
- Combined with global V_{DD} scaling, this method **overcomes the limitations of DVAS**, by contrasting the "Wall-of-slack" phenomenon.
- First ever **application of Back-Biasing to Quality Configurable Systems** (to our knowledge).

Future Developments:

- Devise method for **runtime update** of V_{BB} domains configurations depending on operating conditions (PVT, aging, etc.)
- Investigate **alternative partitioning** techniques (irregular tiling).

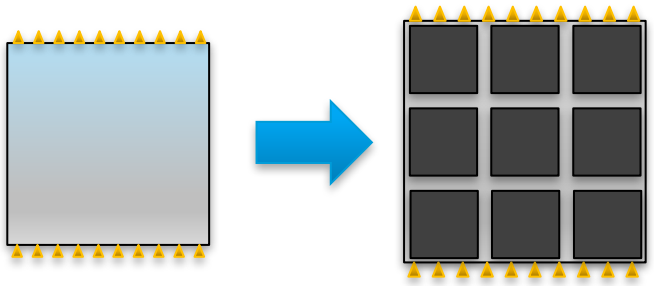

Thank
You
Asante
Shukria
Dhanyavadagalu
Vinaka
köszönöm
Kiitos
Maake
Manana
Dankon
Matondo
Kasih
Terima
Taiku
Diolch i Chi
Tack
Chokrane
Grazie
Arigato
Gracias
Mochchakkeram
Tingki
Paldies
Gratias Tibi
Obrigado
ありがとう
Kia Ora
Kop Khun Khap
Djiere Dieuf
Eskerrik Asko
Go Raibh Maith Agat
ขอขอบคุณคุณ
Najis Tuke
Salamat
Merci
Di Ou Mèsi
Danke
Welalin
Hvala
Nirringrazzjak
Grazas
cảm ơn bạn
धन्यवाद
Biyan
Mauruuru
Kam Sah Hammida
Dank Je
Blagodaram
Ngiyabonga
Dziekuje
Juspaxar
Bedankt
Dàkujem
Ua Tsaug Rau Koj
Děkuji
Suksama
Rahmat
Nuwun
Matur
Misaotra
Danke
Danksheen
спасибо
감사합니다
நன்றி
谢
谢
Хвала

Implementation Flow

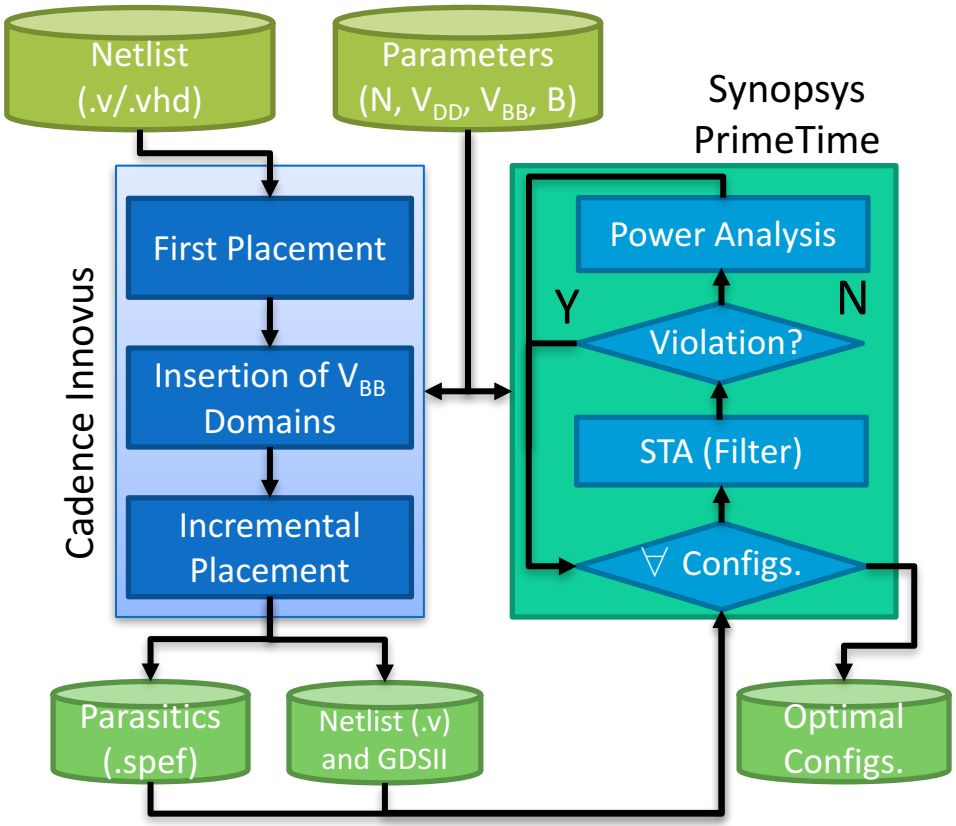


1. Implementation Phase:

- Partition circuit in VBB domains using regular tiling.
- Incremental placement:
  - Insert well-taps
  - Fix possible **constraints violations** due to cell displacement.



Implementation Flow



2. Analysis Phase:

- Exhaustive exploration of all possible configs of Accuracy, V_{BB}, and V_{DD}
- **STA** to prune unfeasible configurations (timing violations)
- **Power analysis** on feasible configs

• Complexity

- Many configurations (thousands), but fast analysis.
- Feasible for **< 10-15** V_{BB} domains