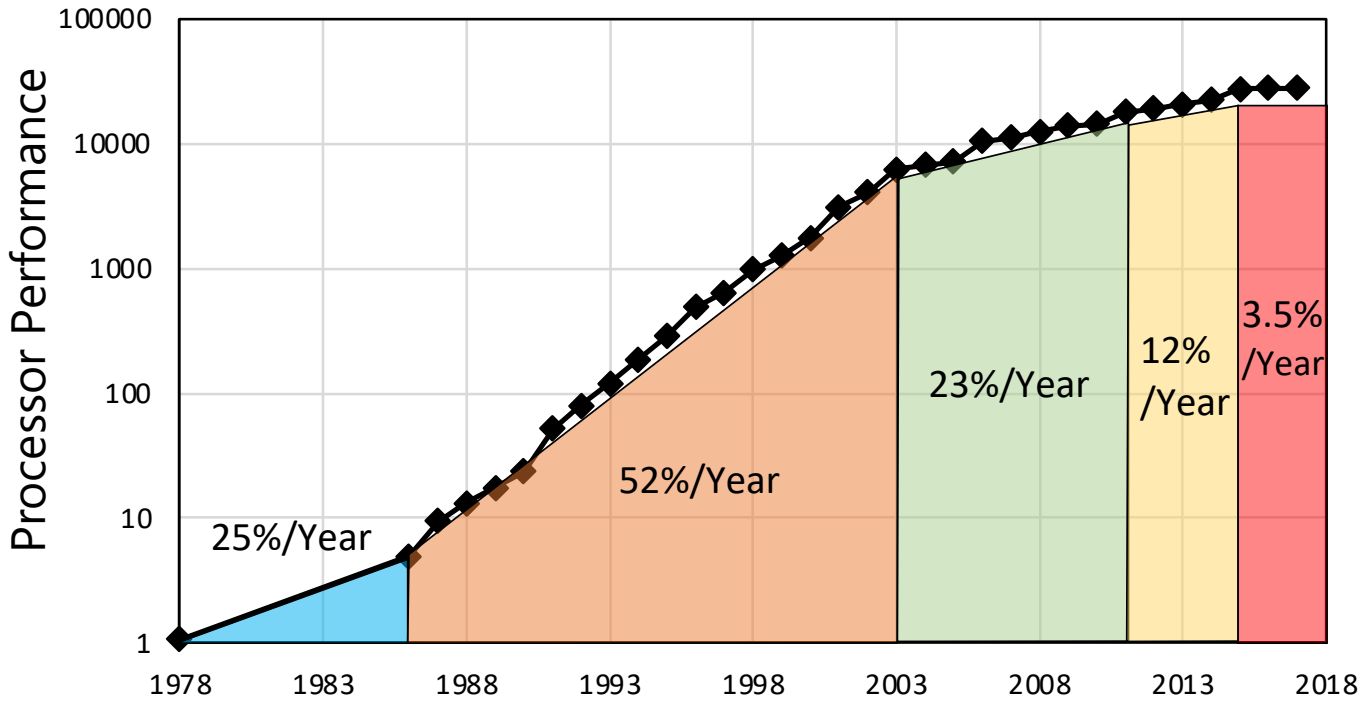
Body Bias Control on a CGRA based on Convex Optimization

Takuya Kojima (UTokyo, Japan),
Hayate Okuhara(NUS, Singapore),
Masaaki Kondo, Hideharu Amano (Keio Univ., Japan)

COOL Chip 25 @Takeda hall, The University of Tokyo, Japan, April 20-22, 2022

A demand for new architectural approaches



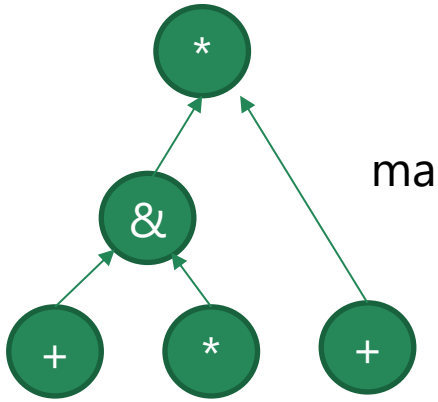
Trend of the processor performance scaling [1]

General-purpose processors are facing a performance improvement limit

- Urgent need for other architectures not depending on the transistor scaling
  - Reconfigurable computing
  - Domain-specific architectures, etc

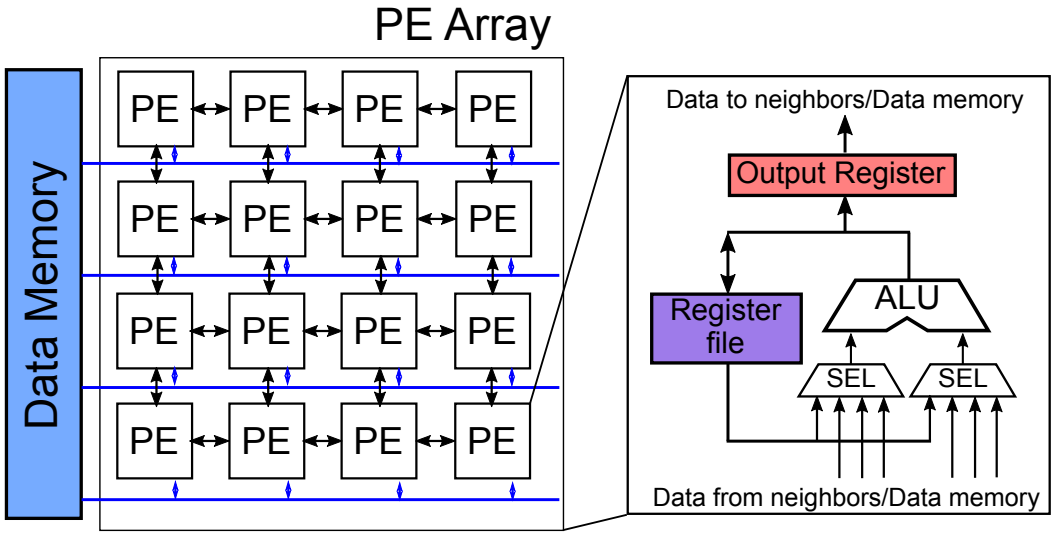
[1] Patterson, D. A., Asanović, K., Hennessy, J. L. (2019). Computer Architecture: A Quantitative Approach.

CGRA: a candidate for future architectures



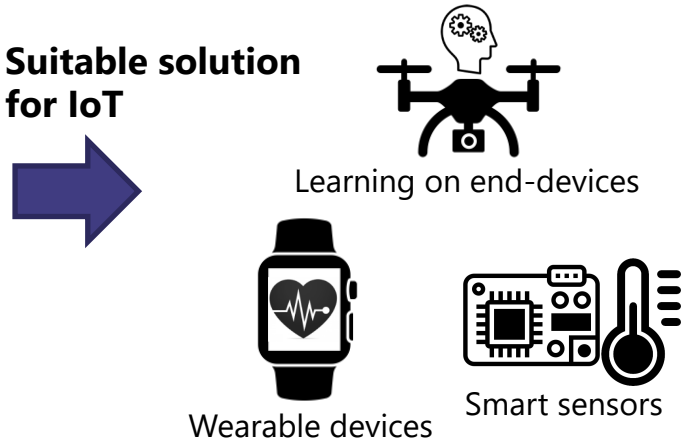
A data flow graph as an application

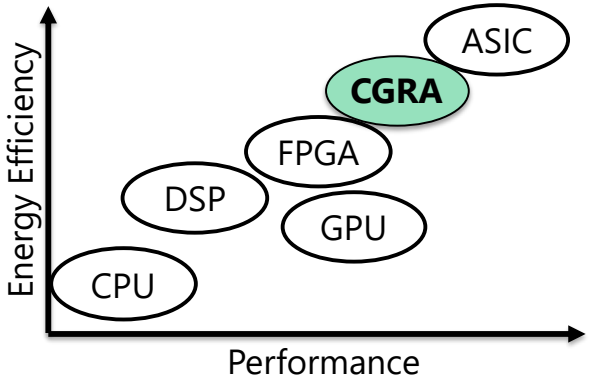
mapping



General structure of the CGRAs

Suitable solution for IoT



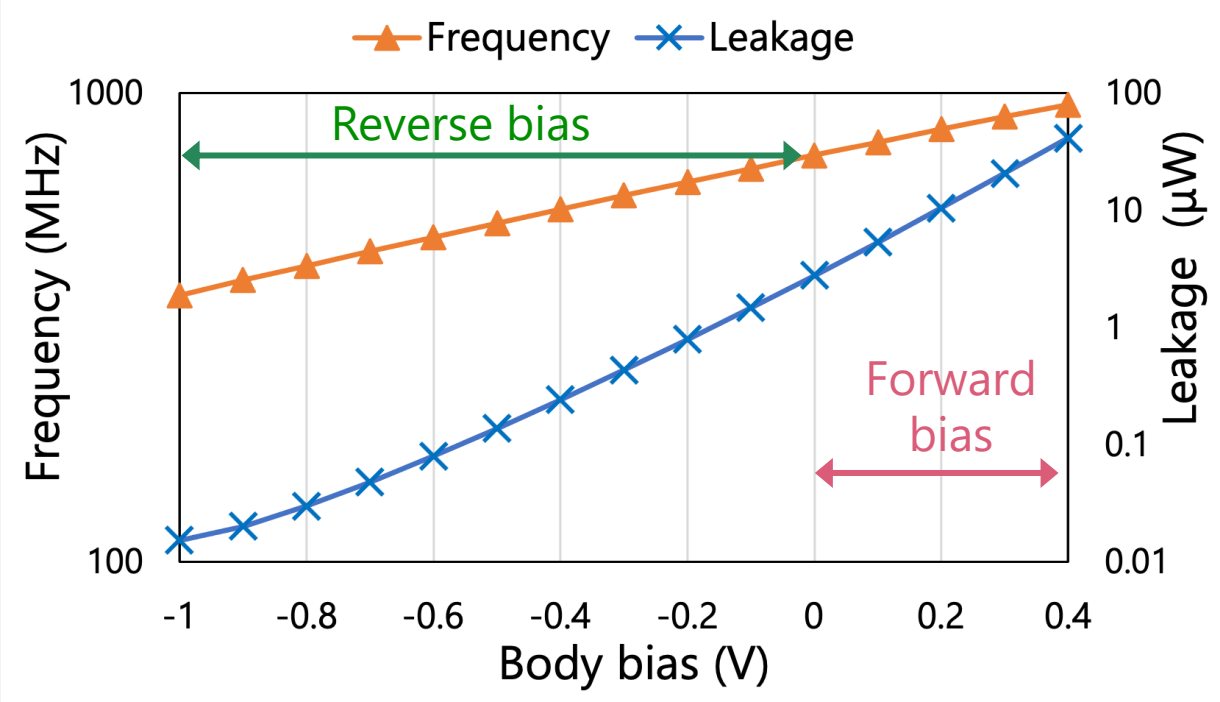


Comparison with other architectures[2]

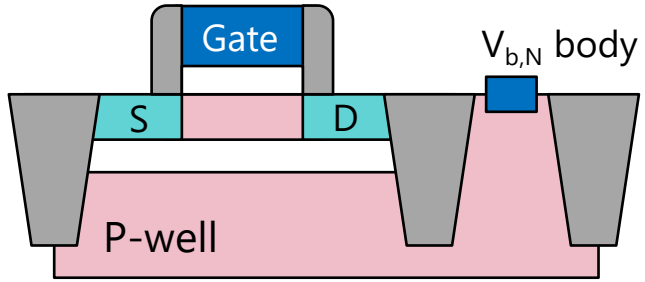
- Coarse-Grained Reconfigurable Architecture (CGRA)
  - Composed of an array of Processing Elements (PEs)
  - Providing a word-level reconfigurability (e.g., 32-bit)
    - Smaller energy-overhead than FPGAs (bit-level)
  - Generally used as an accelerator

[2] Liu, Leibo, et al. "A survey of coarse-grained reconfigurable architecture and design: Taxonomy, challenges, and applications." *ACM Computing Surveys (CSUR)* 52.6 (2019): 1-39.

Body biasing for low-power computing



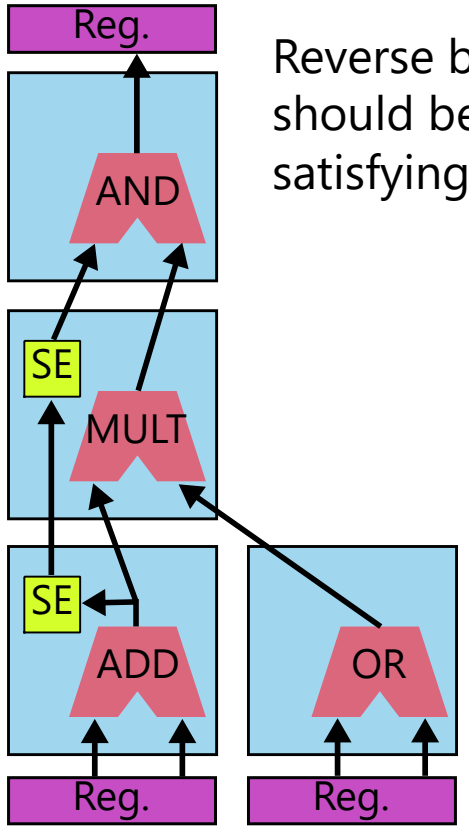
Simulation results of leakage power for a 25-stage ring oscillator composed of FO4 inverters using USJC DDC 55 nm process



N-MOS transistor of an FD-SOI with well contact

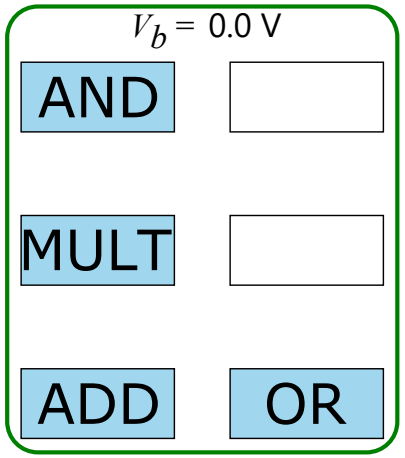
- Body biasing
  - A trade-off b/w performance and leakage power
- With reverse bias (< 0 V)
  - Low performance with Low leakage
- With forward bias (> 0 V)
  - High performance at the cost of leakage

Body bias control on CGRAs

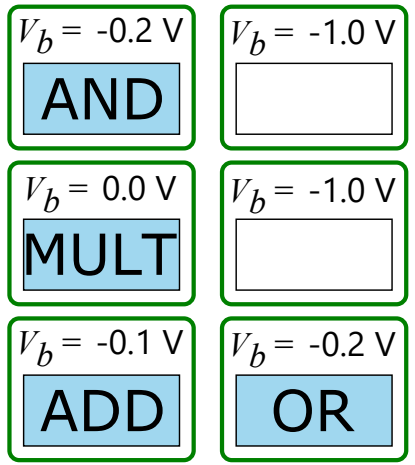


Several data paths configured on the PE array

Reverse bias voltage should be supplied, satisfying the timing constraint



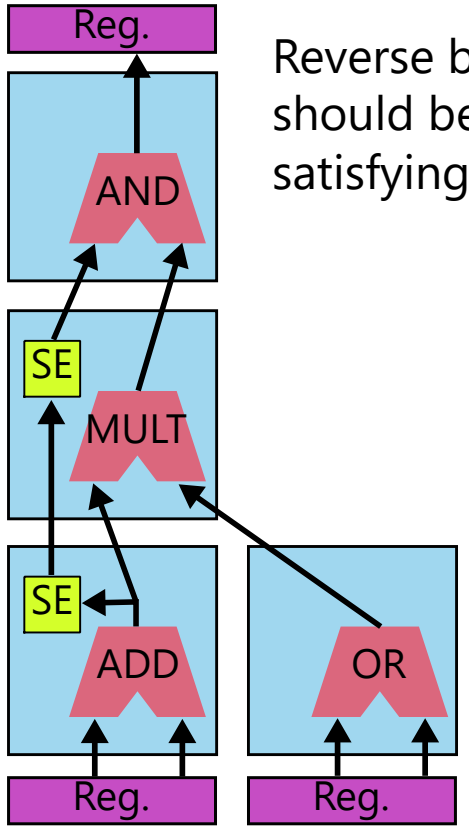
With a single voltage domain



With PE-by-PE voltage domains

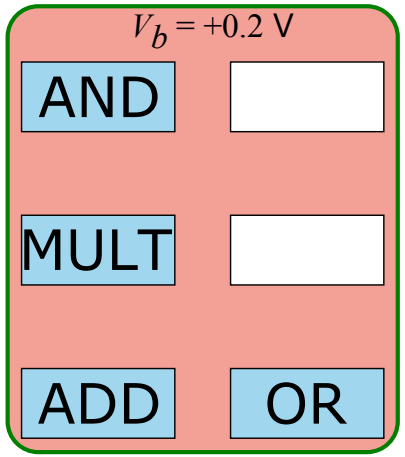
- Fine-grained domains
  - Increases the possibility to utilize the reverse bias to save the leakage power consumption
  - Reduces the cost when forward bias is used

Body bias control on CGRAs



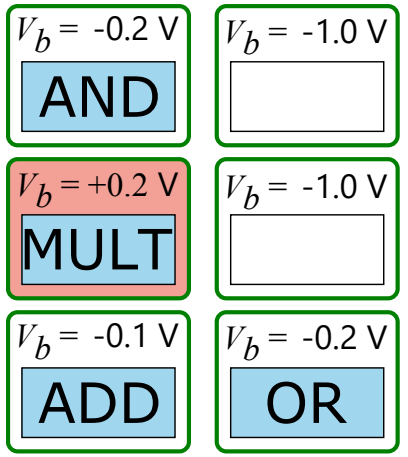
Several data paths configured on the PE array

Reverse bias voltage should be supplied, satisfying the timing constraint



With a single voltage domain

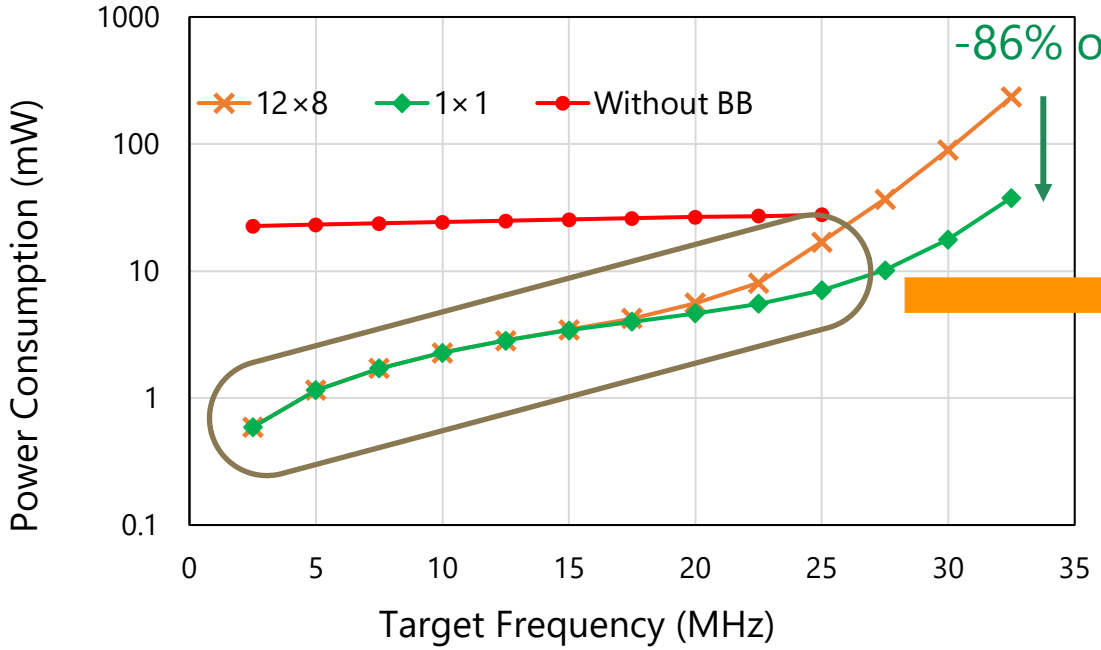
←Incurring a prohibitive high increase of leakage

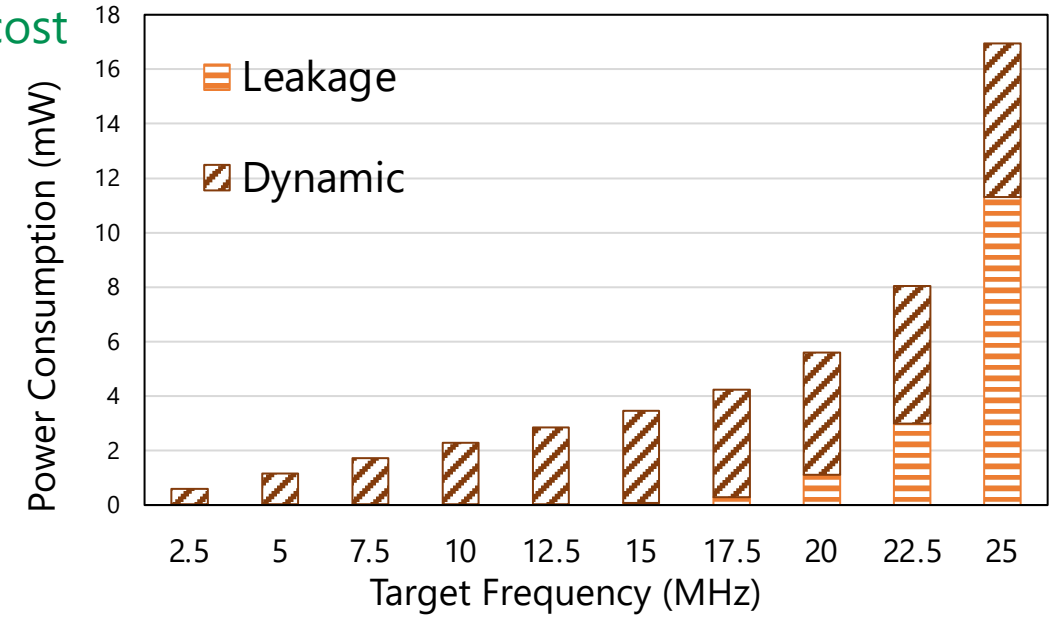


With PE-by-PE voltage domains

- Fine-grained domains
  - Increases the possibility to utilize the reverse bias to save the leakage power consumption
  - Mitigates the penalty cost when forward bias is used

Impact of body bias control on a CGRA





12x8 domain size case (< 25MHz)

- A preliminary analysis based on a CGRA shows
  - Reduction of power consumption adaptively
  - Performance enhancement by forward bias
  - Minimized leakage cost of forward bias by the fine-grained domain partitioning

Technical challenge

Given an operational frequency as the timing constraint (D_{req}),
the CGRA compiler has to determine the voltages to minimize the leakage

- The problem is defined as follows:

$$\min P_{\text{leak}} = \sum_{i=0}^{N_{\text{dom}}-1} P_{\text{leak},i}(V_{b,i})$$

Too complicated to solve the problem

Because of

$$I_{\text{sub}} = I_{\text{off}} 10^{\frac{V_{gs}+\eta(V_{ds}-V_{DD})-k_\gamma V_{sb}}{S}} (1 - e^{\frac{-V_{ds}}{v_T}})$$

Subthreshold leakage current [3]

subject to

$$\forall D_l < D_{\text{req}} \ (0 \leq l < N_{dp})$$

$$D_l = \sum_{v \in l\text{-th datapath}} D_v(V_b)$$

[3] Weste, Neil HE, and David Harris. *CMOS VLSI design: a circuits and systems perspective*, 2015.

Optimality and scalability issues in prior work

- An approach based on genetic algorithm [4]
  -  Tolerant to large scale problem (i.e., with fine-grained domains)
  -  Impossible to guarantee the optimality
  -  Long time to find a solution
- Another approach based on Integer Linear Program (ILP) [5]
  -  Always providing the optimal solution
  -  Less scalability due to the NP-completeness of ILP

[4] Matsushita, Yusuke, et al. "Body bias grain size exploration for a coarse grained reconfigurable accelerator." *2016 26th International Conference on Field Programmable Logic and Applications (FPL)*. IEEE, 2016.
[5] Kojima, Takuya, et al. "Body bias optimization for variable pipelined CGRA." *2017 27th International Conference on Field Programmable Logic and Applications (FPL)*. IEEE, 2017.

ILP-based method [5]

- Considering discreate voltages, the problem is formulated as follows:

Binary decision variable

$$isV_{b,ij} = \begin{cases} 1 & \text{if the } i\text{-th domain} \\ & \quad \text{is set with } j\text{-th } V_b \\ 0 & \text{otherwise} \end{cases}$$

Objective function

$$\min P_{\text{leak}} = \sum_{i=0}^{N_{\text{dom}}-1} \sum_{j=0}^{N_{\text{bb}}-1} P_{leak,i,j}\ isV_{b,ij}$$

An example of the leakage table $P_{leak,i,j}$

j	V_b	Leakage power of domain 0 (i=0)
0	-0.8	0.197 uW
1	-0.6	0.236 uW
…	…	
6	+ 0.4	7.89 uW

Constraints

$$\sum_{j=0}^{N_{\text{bb}}-1} isV_{b,ij} = 1 \quad \forall j = \{0, 1, \ldots, N_{\text{dom}} - 1\}$$

$$\forall D_l < D_{\text{req}}\ (0 \leq l < N_{dp})$$

$$D_l = \sum_{v \in \substack{l\text{-th} \\ \text{datapath}}} \sum_{j=0}^{N_{\text{bb}}-1} D_{v,j}\ isV_{b,ij}$$

Towards scalable method

- To address the scalability issue of the ILP-based method, this work tries to reformulate the problem as a convex optimization
  - Convex optimization
    - Objective function and all the constraints are described as convex functions
    - Polynomial time algorithms (e.g., [6]) are available even for non-linear functions
- An approximate model of the subthreshold leakage [7] is used

$$I_{\mathrm{leak}} = I_{\mathrm{leak0}} \exp(AV_{DD} + BV_b + CT)$$

- Delay time for each component is calculated with α-power law [8]

$$\tau = k\frac{CV_{DD}}{(V_{DD} - V_t)^{\alpha}} \qquad V_t = V_{t0} - K_{\gamma}V_b$$

[6] Andersen, Erling D.,et al. "On implementing a primal-dual interior-point method for conic quadratic optimization." *Mathematical Programming* 95.2 (2003): 249-277.
[7] Fujita, Yu, et al. "Power optimization considering the chip temperature of low power reconfigurable accelerator CMA-SOTB." *2015 Third International Symposium on Computing and Networking (CANDAR)*. IEEE, 2015.
[8] Sakurai, Takayasu, and A. Richard Newton. "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas." *IEEE Journal of solid-state circuits* 25.2 (1990): 584-594.

Formulation for the convex optimization

- The standard form of convex optimization

Objective function

$$\min f_0(\boldsymbol{x})$$

Constraints

$$f_i(\boldsymbol{x}) \leq 0, i \in 1, ..., m$$

- In this work,
  - A vector $\boldsymbol{x}$: a set of body bias voltages

$$\boldsymbol{x} = [V_{b,0}, V_{b,1}, ..., V_{b,N_{\mathrm{dom}}-1}]$$

$$\forall i \; V_{b,\mathrm{lbound}} \leq V_{b,i} \leq V_{b,\mathrm{ubound}}$$

  - Objective function

$$\min P_{\mathrm{leak}}(\boldsymbol{x}) = \sum_{i=0}^{N_{\mathrm{dom}}-1} I_{\mathrm{leak0},i} \exp(AV_{DD} + BV_{b,i} + CT)$$

$$I_{\mathrm{leak0},i} = N_{\mathrm{PE},i} \times I_{\mathrm{leak0,PE}}$$

  - Constraints: $\forall l \boldsymbol{D_{0,l}} \boldsymbol{s}^{\mathrm{T}} \leq D_{\mathrm{req}} \; (0 \leq l < N_{dp})$
    - Total delay time of path l with zero bias

$$\boldsymbol{D_{0,l}} = [D_{0,l0}, D_{0,l1}, ..., D_{0,lN_{\mathrm{dom}}-1}]$$

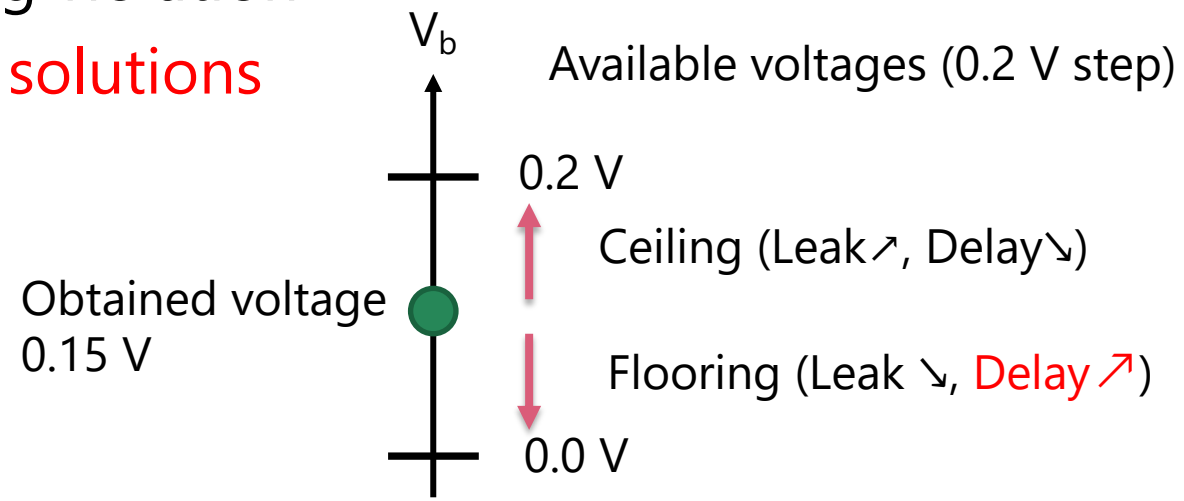
    - Delay scale

$$\begin{aligned} \boldsymbol{s} &= S(\boldsymbol{x}) \\ &= [S(V_{b,0}), S(V_{b,0}), ..., S(V_{N_{\mathrm{dom}}-1})] \\ S(V_b) &= \frac{(V_{DD} - V_{t0})^{\alpha}}{(V_{DD} - V_{t0} + K_{\gamma} V_b)^{\alpha}} \end{aligned}$$

Voltage rounding strategies

- Given that the available body bias voltages are discrete, the voltages have to be rounded to the available voltages
- The most straightforward way of rounding

  → All voltages are ceiled
  because only flooring could occur a timing violation
  - However, it would miss smaller leakage solutions
- Two strategies are proposed
  1. Heuristic with $\mathcal{O}(N_{\mathrm{dom}})$ complexity
     - Allowing non optimal rounding
  2. Exact rounding based on an ILP
     - A case with two voltages in the ILP formulation (i.e., N_{bb}=2)



Flow of rounding heuristic

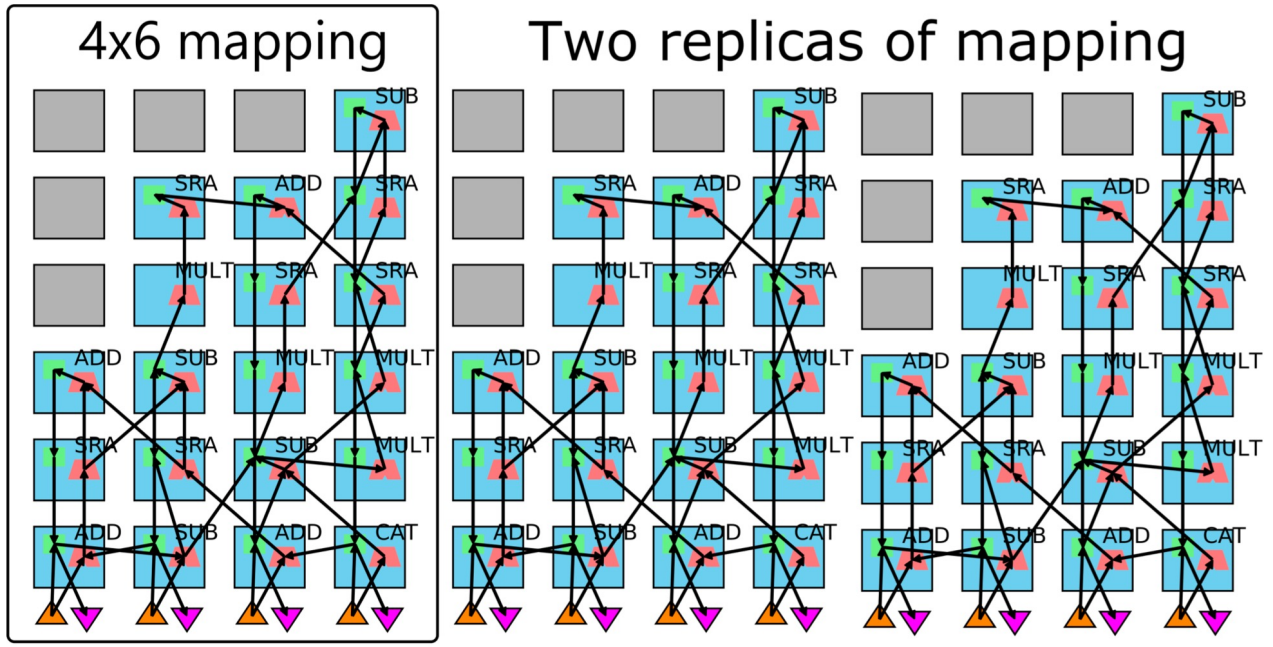
Input: the solution of convex programming $\boldsymbol{x}$
Output: the rounded voltages $\boldsymbol{X}$

1: $\boldsymbol{X} \leftarrow \text{Floor}(\boldsymbol{x})$ /* *Firstly, all of them are floored* */
2: $U \leftarrow \text{sorted\_index}(\boldsymbol{x})$
/* *Asc. order of leakage increase by ceiling* */
3: **while** $U \neq \emptyset$ and isTimingViolate($\boldsymbol{X}$) **do**
4: Get an index i of the first element in U
5: $\boldsymbol{X}[i] \leftarrow \text{Ceil}(\boldsymbol{x}[i])$
6: $U \leftarrow U - \{i\}$ /* *Eliminating the 1st element* */
7: **end while**

- It tries to floor all the voltage at the beginning
- Then, repeat ceiling one voltage until the timing constraint is met
  - The order of ceiled voltages is the ascending order of leakage increase by ceiling
  - i.e., The voltage occurring the smallest increase is firstly ceiled

Experimental setup-1

- A studied CGRA: VPCMA2 [8]
  - The PE array size: 8 x 12
- 7 Benchmark appellations
  - Image processing
    - Gray scale, 8bit sepia filter, 24bit sepia filter, alpha blender
  - Signal processing
    - 4-point DCT, 4-point FFT
  - Encryption
    - AES



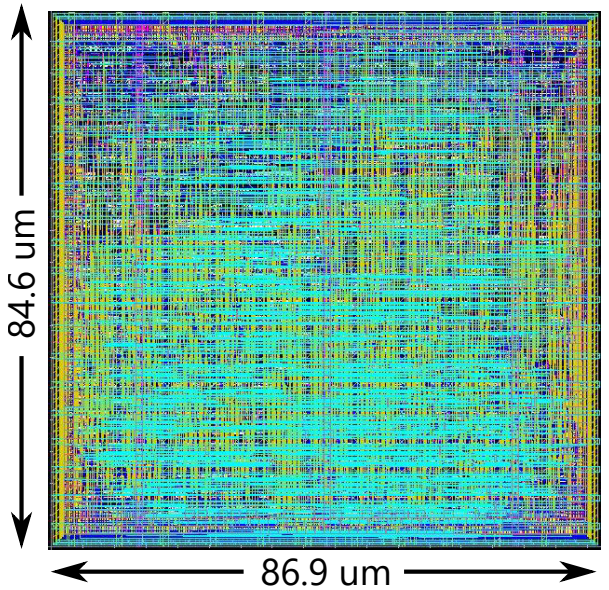
4x6 DCT mapping with two replicas

Configurable pipelined registers are omitted

- Different sizes of mappings are prepared for each appellations

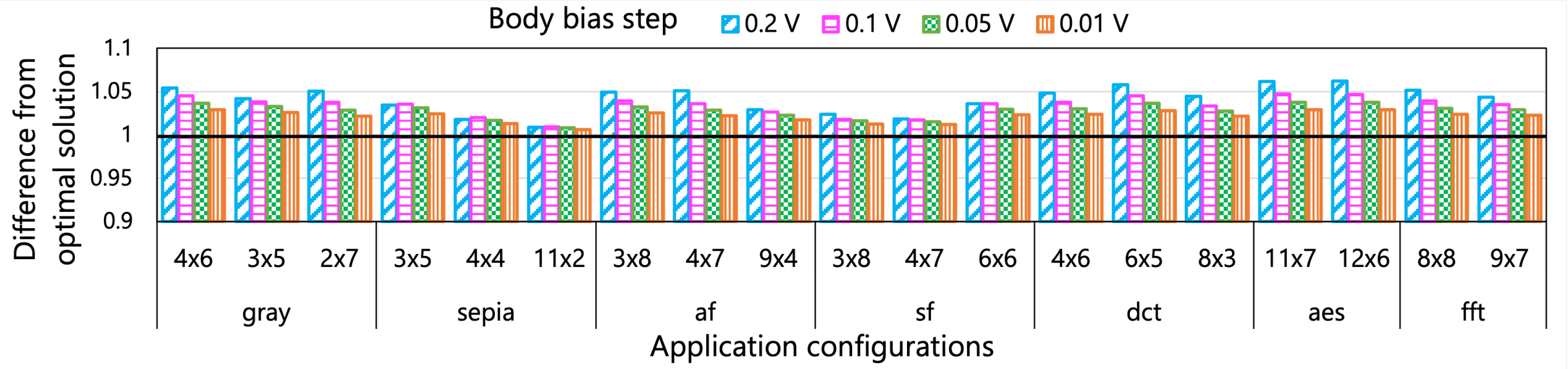
Experimental setup-2

- Implementation to obtain the parameters
  - Process: USJC 55nm DDC
  - Synthesis: Synopsys design compiler
  - Layout: Cadence Innorvus
  - Leakage and delay time: Synopsys HSPICE
- Voltage conditions
  - Resolution : 0.2 V, 0.1 V, 0.05 V, 0.01 V
  - Range: -0.8 V –+0.2 V
- Optimization software executed on Ryzen Threadripper 3960X
  - ILP: Gurobi (solver), PuLP (modeler)
  - Convex optimization: mosek (solver), CVXPY (modeler)



The PE layout

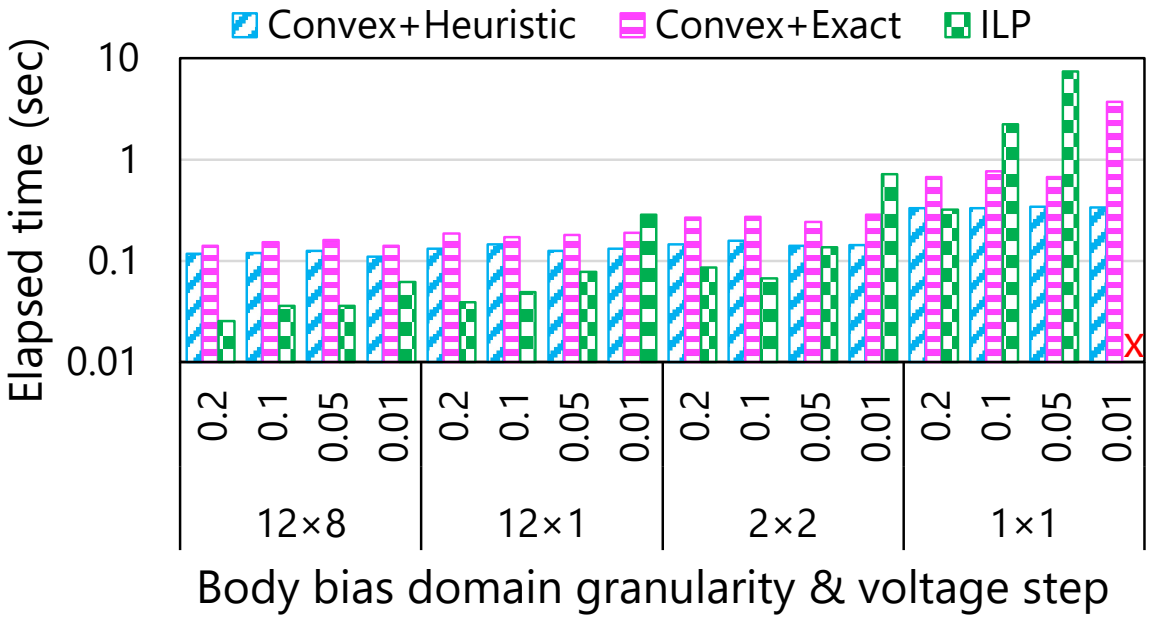
Optimality gap analysis



Normalized differences in the optimization results between the proposed method (rounding heuristic) and the ILP-based method

- The cases for 0.2 step shows larger errors compared to the other steps
- For 0.1 V step or finer resolutions, the error is less than 5%
- The results with the exact rounding includes around 0.1% error

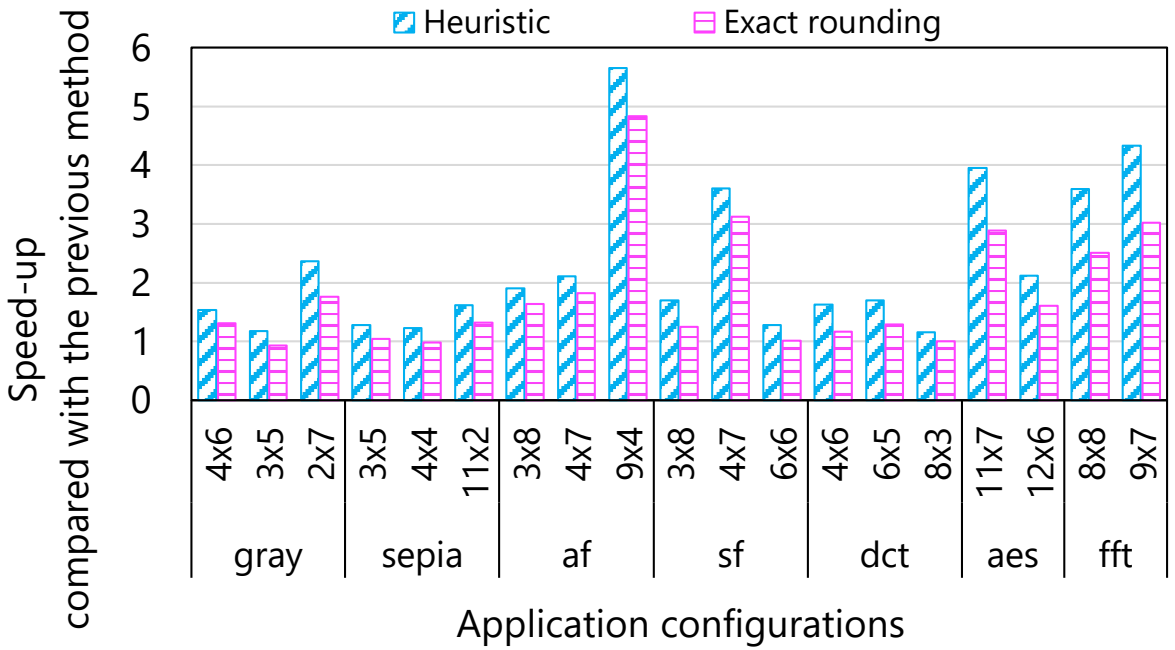
Elapsed time comparison



Elapsed time for each method when *FFT* is mapped to 9x7 PEs
(Only four interesting domain granularities are shown)

- When the solution space is small (e.g., a single domain, 12x8),
  - The ILP method is faster
- In the case of the biggest problem
  - 0.01 V step and 1x1 grain size
    → 141 voltage candidates for 96 domains
  - The ILP cannot be solved in 3 hours
  - In contrast,
    the proposed methods take 0.45 sec and 4.2 sec, respectively with the heuristic and ILP-based rounding

Speedup of the proposed methods



Speed-up compared to the ILP-based method
for a middle-class problem (3x2 grain size & 0.1 V step)

- With the rounding heuristic
  - 2.32x speed-up, on average
- With the exact rounding
  - 1.85x speed-up, on average
  - But longer time for some cases (e.g., *gray* 3x5 mapping)

Conclusion

- A scalable body bias optimization method for CGRAs was proposed
  - By introducing an approximated leakage model and
  - By reformulating the problem as a convex optimization problem
- In addition, two rounding methods were presented
- Evaluation results demonstrated
  - The optimization results with the proposed method contain a negligible error (< 5% for 0.1 V or finer voltage resolution)
  - Compared to the previous method based on an ILP,
    the proposed method can solve the problem 2.32x faster, even for a middle class of the problem
  - The proposed method can quickly solve the biggest problem, which cannot be solved by the previous one