

1EOr2B-02

Lowering Latency in High-Speed Gate-Level-Pipelined Single Flux Quantum Datapath Using Interleaved Register File

Akira Fujimaki

Collaborators

Ryota Kashima, Ikki Nagaoka, Tomoki Nakano

Masamitsu Tanaka, Taro Yamashita

Nagoya University

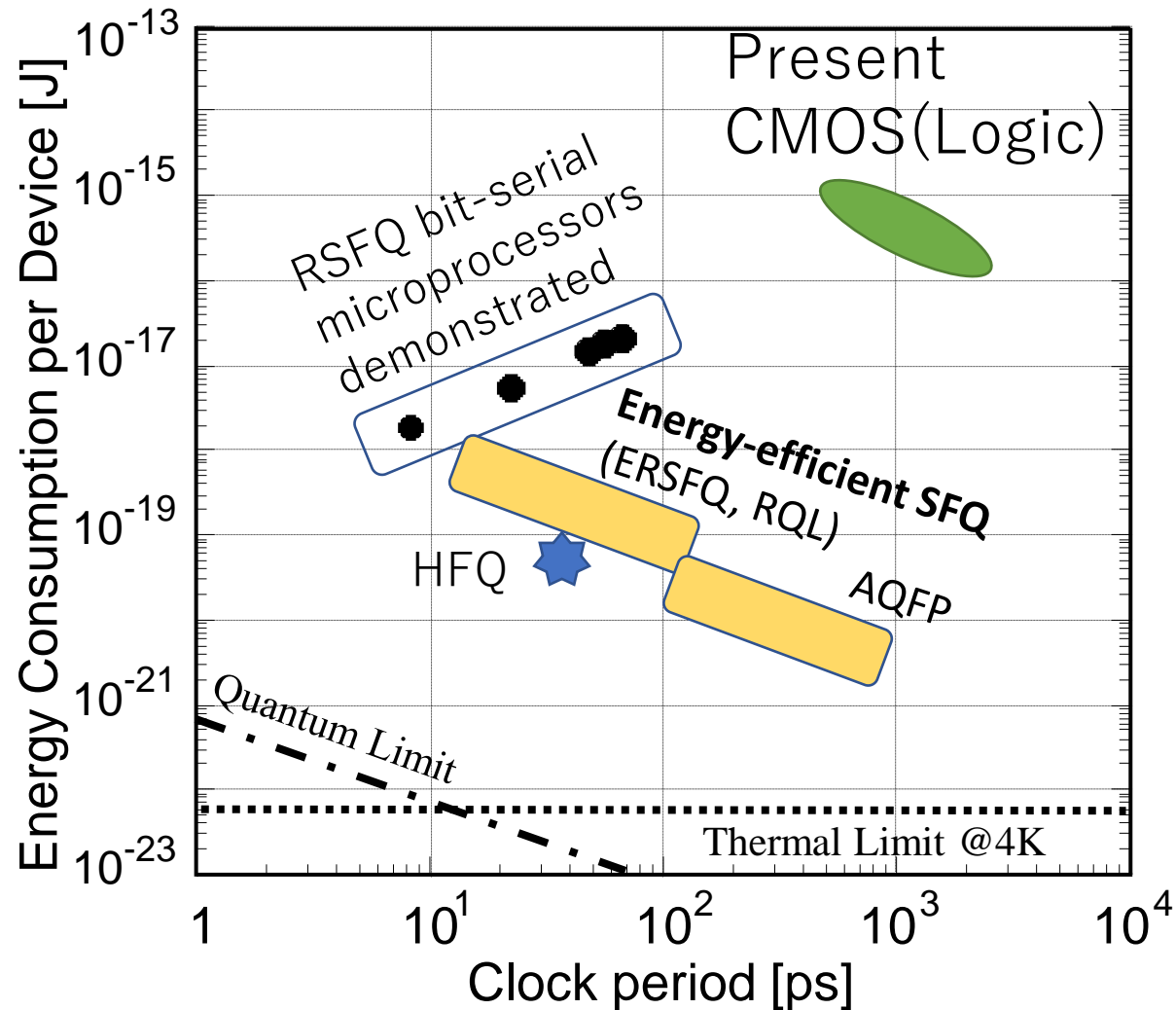
Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP18H05211, JP19H01105, and JP18H01498 and by VDEC of the University of Tokyo in collaboration with Cadence Design Systems, Inc. The test circuits except π junctions were fabricated at CRAVITY, AIST.

Outline

- Superconductor-based supercomputer in future
- High-speed matrix memory
- High-speed processors
 - Way to high-speed operation
 - Low latency in high-speed gate-level-pipelined SFQ datapath
 - Other SFQ processors
- Summary

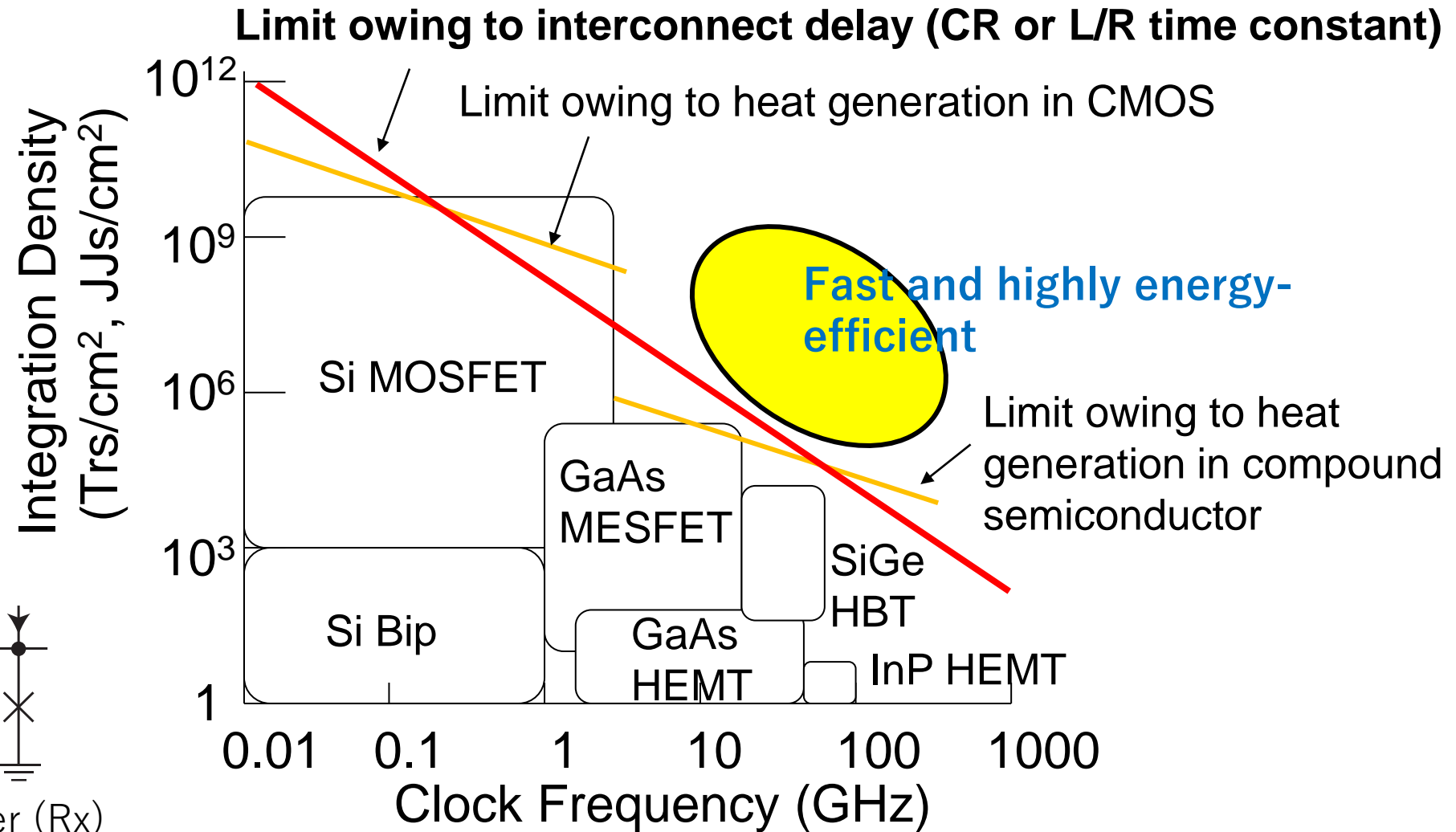
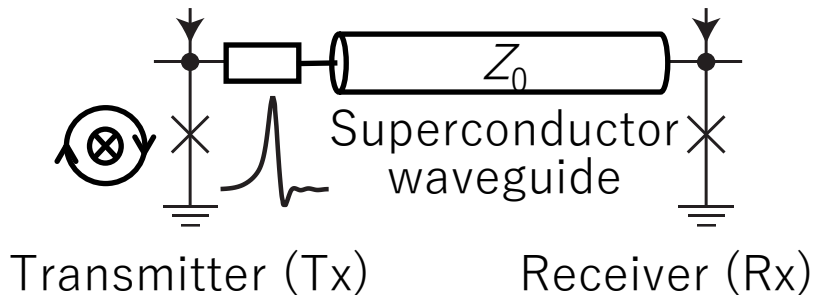
Energy-Efficiency



Presentations about HFQ circuit

- Oral presentations
 - 3ESpeOr4-03 D. Pham
 - 4EOr2A-02 T. Yamashita
- Poster presentations
 - 2EPo1A-03 F. Li
 - 3EPo2B-02 M. Higashi
 - 3EPo2B-08 S. Tanemura

Shortened Interconnect Delay with PTLs



Superconductor Digital Electronics in Transition

Dr. Mukhanov's plenary
talk at **ASC 2006**

- Applications
 - Small-scale applications such as Digital-RF receivers
- Fab
 - Reliable but integration scale is low
- Logic/Memory
 - Fast RSFQ logic, but no large random access memory (RAM)

Dr. Mukhanov's plenary
talk at **ASC 2014**

- Applications
 - Energy-efficient high-end computing
- Fab
- Logic/Memory
 - Energy-efficient logic
 - Several proposals of cryo-memories

ASC 2022

- Applications
 - High-end computing **with quantum computers**
- Fab
 - **π -junctions** are available
- Logic/Memory
 - **Ultra high-speed, bit-parallel logic**
 - **High-speed matrix memory**

Image of a Superconductor-Based Energy-Efficient Supercomputer



Dilution Refrigerator

General Purpose Computing @4 K

- Classical computing, but high-speed and low-power
- A single flux quantum is used as an information carrier, *i.e.*, Single Flux Quantum (SFQ) circuit

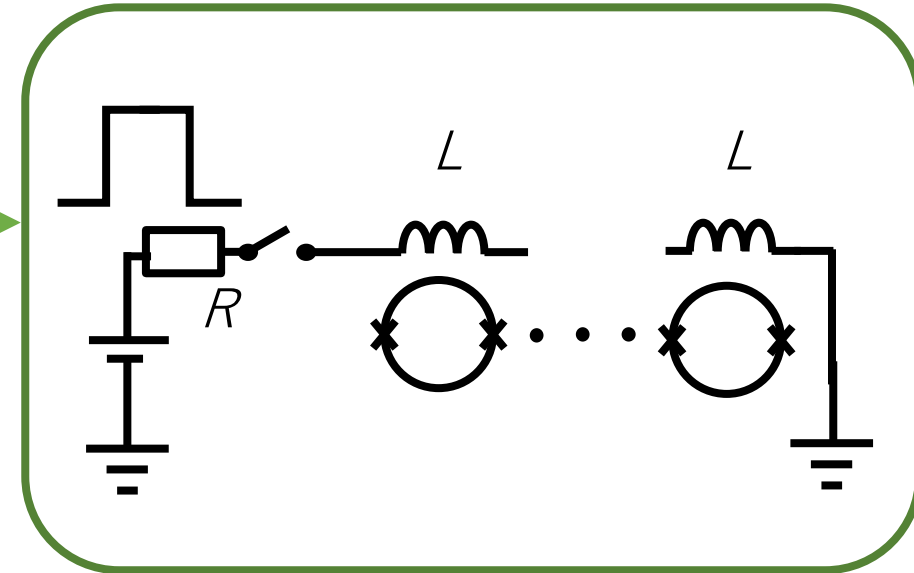
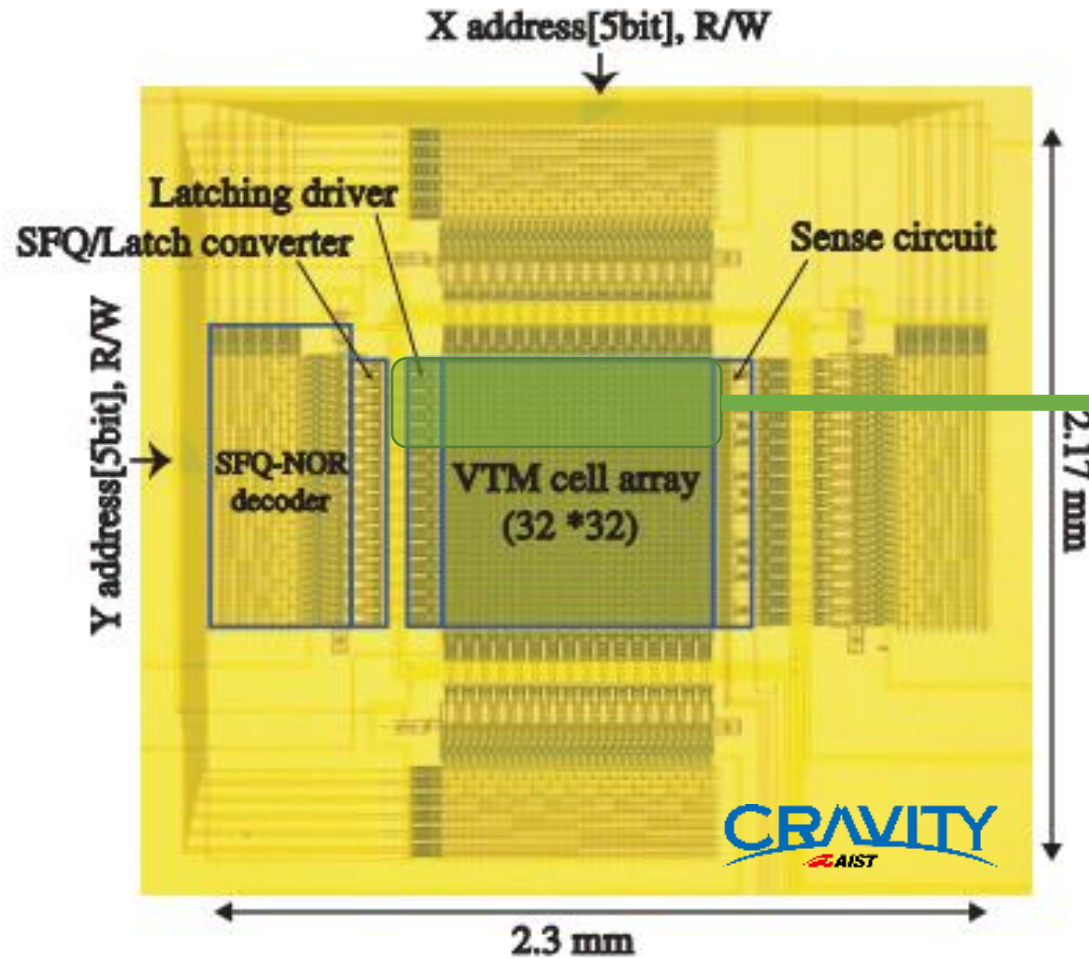
Special Purpose Computing @10 mK

- **Quantum Gate Computing**
 - Extremely fast for several-types of problems
 - Classical digital and analog circuits, *i.e.*, an oscillator, detectors, ADCs, DACs, DSPs are needed for executing quantum computing
- Quantum Annealing based on tunneling
 - Suitable for combinatorial optimization
 - Classical computing is needed for initialization

Outline

- Superconductor-based supercomputer in future
- **High-speed matrix memory**
- High-speed processor
 - Way to high-speed operation
 - Low latency in high-speed gate-level-pipelined SFQ datapath
 - Other SFQ microprocessor
- Summary

Fundamental Wall to be Overcome



Recharge process makes power consumption higher and limits high-speed operation.

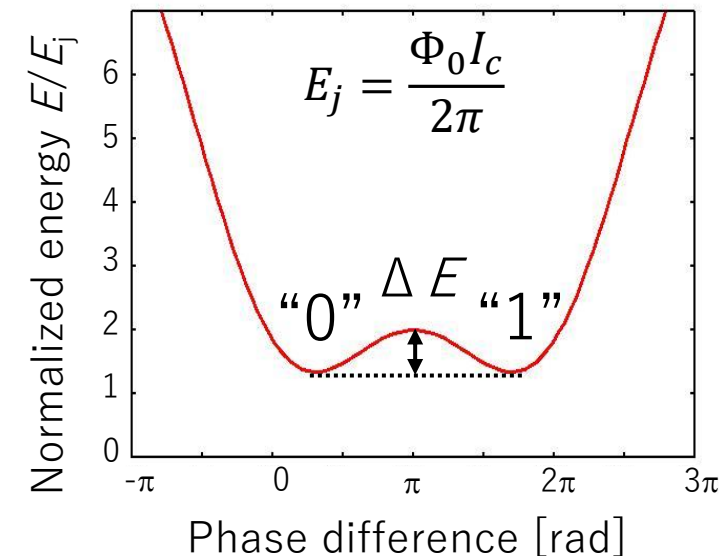
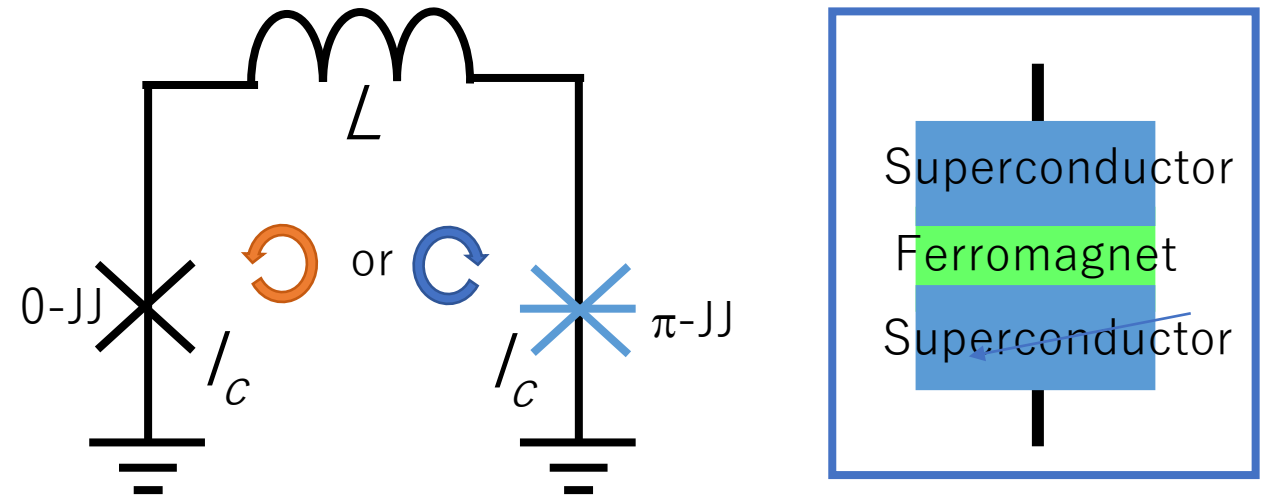
Impulse-driven memory is needed

Requirements for Impulse-driven memory

1. Double well potential is required.
Two potential minima corresponding to a binary bit are needed.
2. The barrier height ΔE needs to be small.
For writing data, the ΔE needs to be smaller than an energy of the impulse signal.
3. The ΔE needs to be tunable.
The operation needs to be performed only for a selected memory cell.

$$\Delta E < E_{\text{write}} < \Delta E'$$

$$E_{\text{pulse}} \approx 10^{-21} \sim 10^{-20} \text{ J}$$



Development of High-Speed Impulse-Driven Memory

Demonstrated

Proposal of impulse driven memory [1]

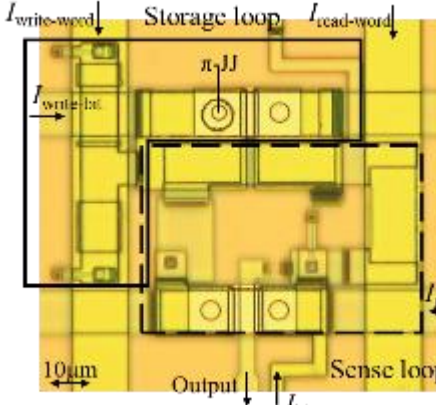
Demonstration of full-operation of 1bit memory cell [2]

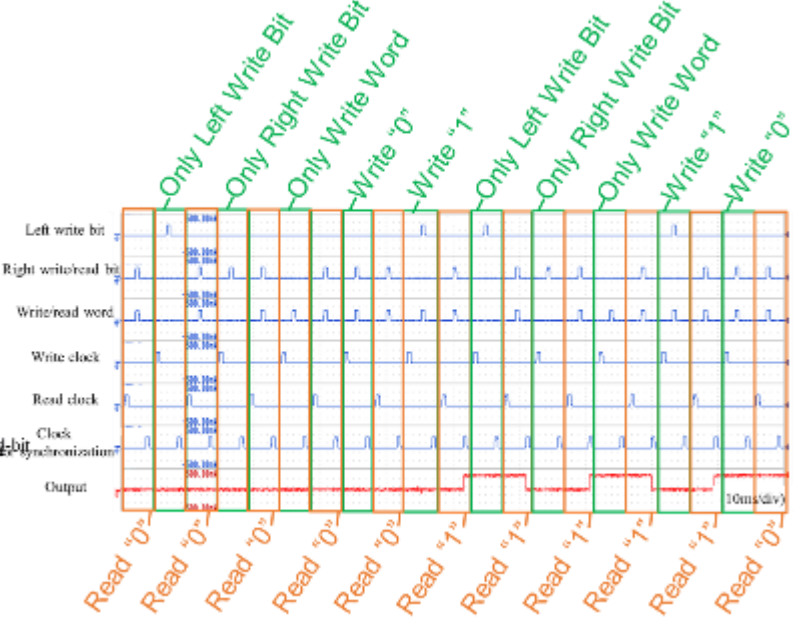
Demonstration of multi-bit array memory

Now Tackling

Demonstration of a large-scale matrix memory

Impulse-driven memory cell





- Write/Read-operations of 1bit memory cell is demonstrated only with impulses signals.

[1] Y. Takeshita *et al.*, *IEEE Trans. Appl. Supercond.*, 31, 5 (2021).

[2] H. Fujisawa, Y. Takeshita *et al.*, SSV2022, P-21, Kyoto, Sep. 2022.

Outline

- Superconductor-based supercomputer in future
- High-speed matrix memory
- High-speed processors
 - Way to high-speed operation
 - Low latency in high-speed gate-level-pipelined SFQ datapath
 - Other SFQ processors
- Summary

Design Fundamentals

- **Timing deviation from designed values @4K**
 - **Jitter** : **0.065 ps** for a JJ in JTL, **0.36ps** for a balanced comparator
(Measured with a high-resolution TDC)
 - **Chip-to-chip variation (Global spread): 0.2 ps/JJ**
- **Flexible PTL technologies available for 50 G bps**
 - Develop multicast PTL-drivers
 - up to **4-fanouts**
 - Optimize via-contact structure
 - up to **54 via-contacts** in a PTL

No physical restrictions for 50-GHz-operations even in bit-parallel processing

➡ Designing techniques should be improved.

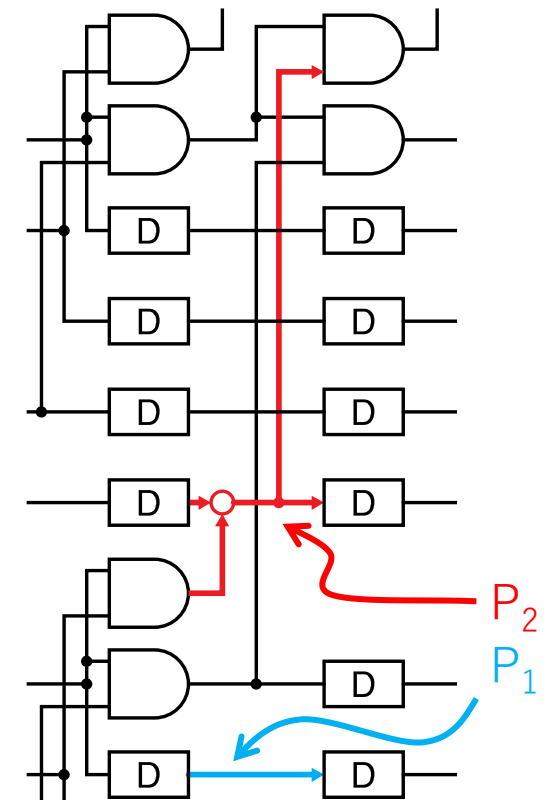
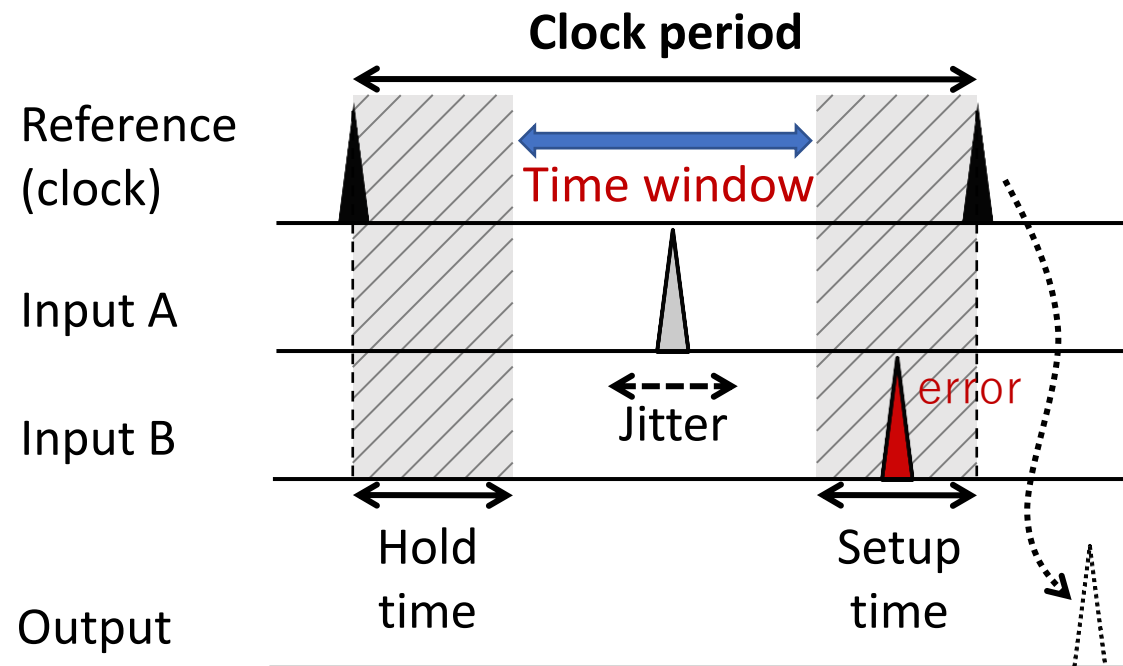
Timing Design for Parallel Processing

Difficulty

- Each gate has different hold time and setup time.

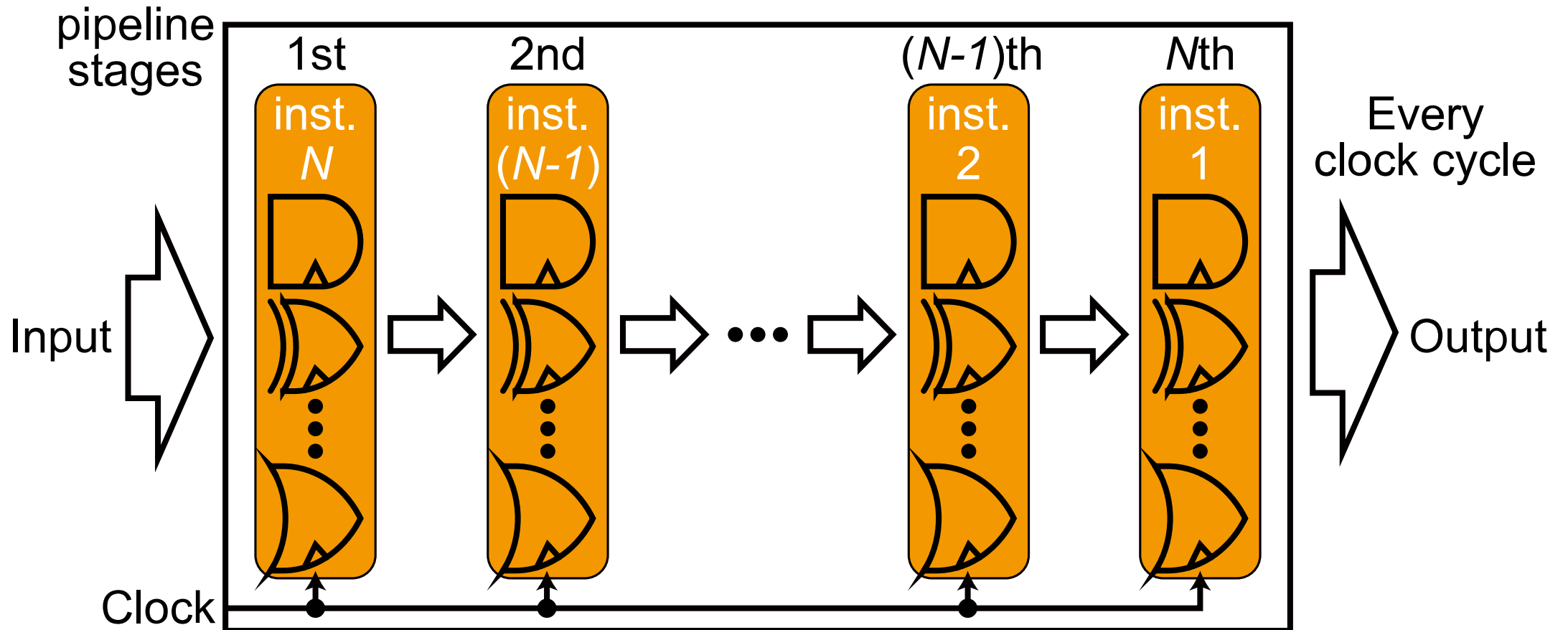
Present Approach

- The timing of input pulses is adjusted so as to be set at the center of the time window.



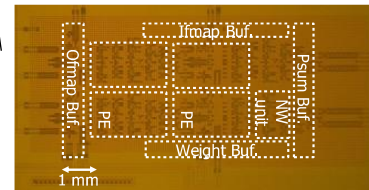
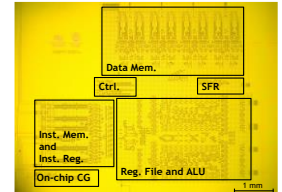
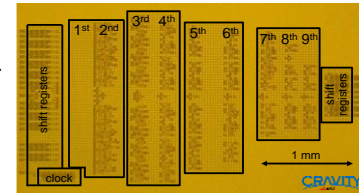
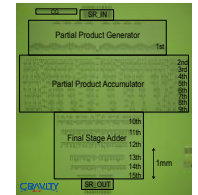
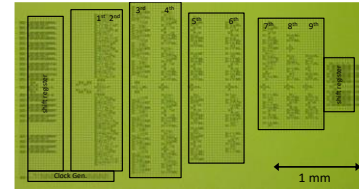
Gate Level Pipelining

- Frequency can be maximized.
- No additional pipeline registers are needed owing to the inherent latch function.



Demonstrated Gate-Level-Pipelined Circuits

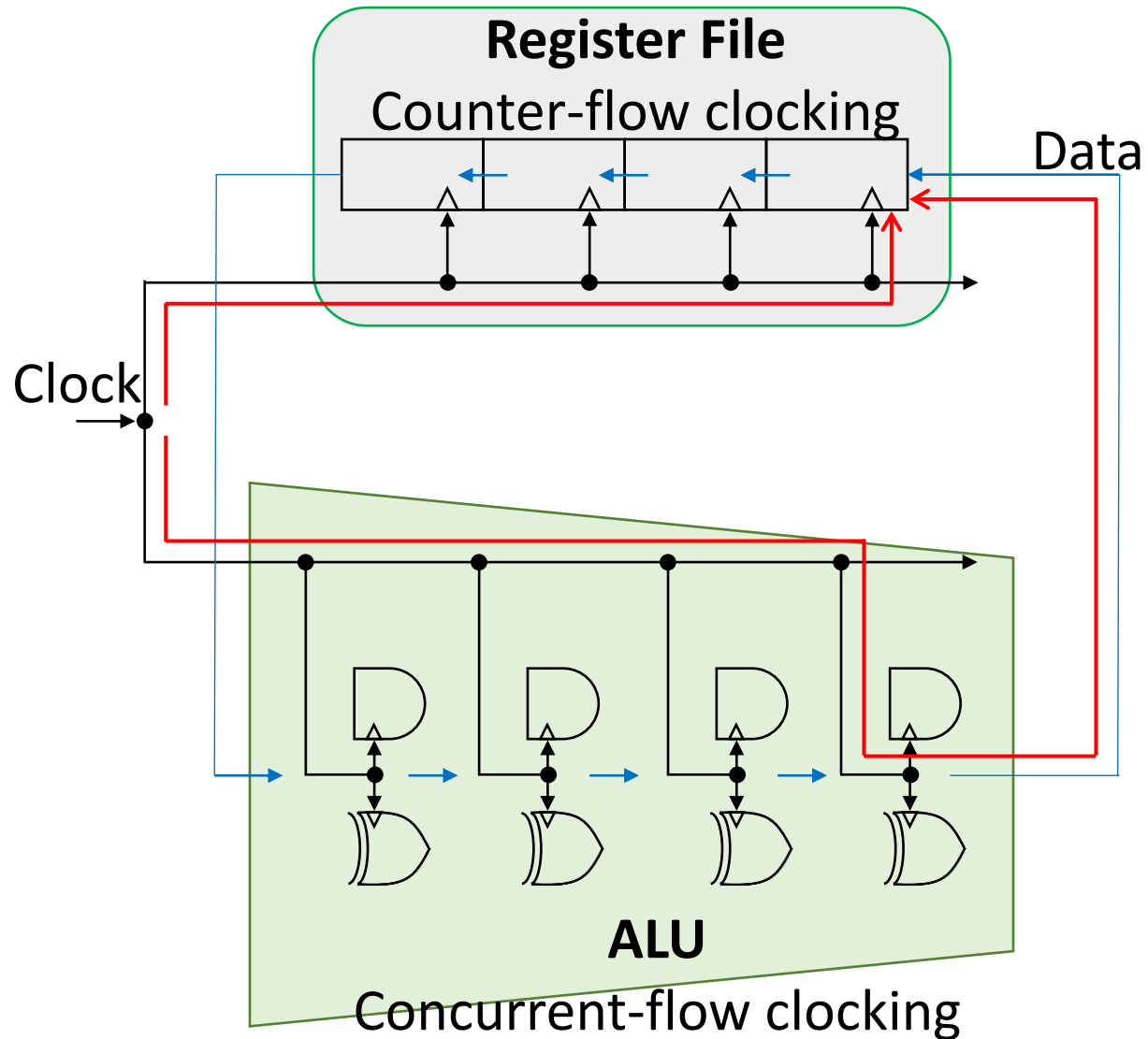
Fabricated Chip	Frequency (GHz)	Power (mW)	Efficiency (TOPS/W)	JJ Count
(a) 8-bit ALU	56	1.6	40	4,846
(b) 8-bit array multiplier	48	5.6	8.5	20,251
(c) 8-bit low-voltage ALU	30	0.28	109	7,451
(d) 4-bit low-voltage multiplier	52	0.13	381	4,498
(e) 4-bit microprocessor	32	6.5	2.5	25,403
(f) 2x2 systolic PE array	34	0.71	382	9,263



Challenges for high-speed SFQ GLP circuits

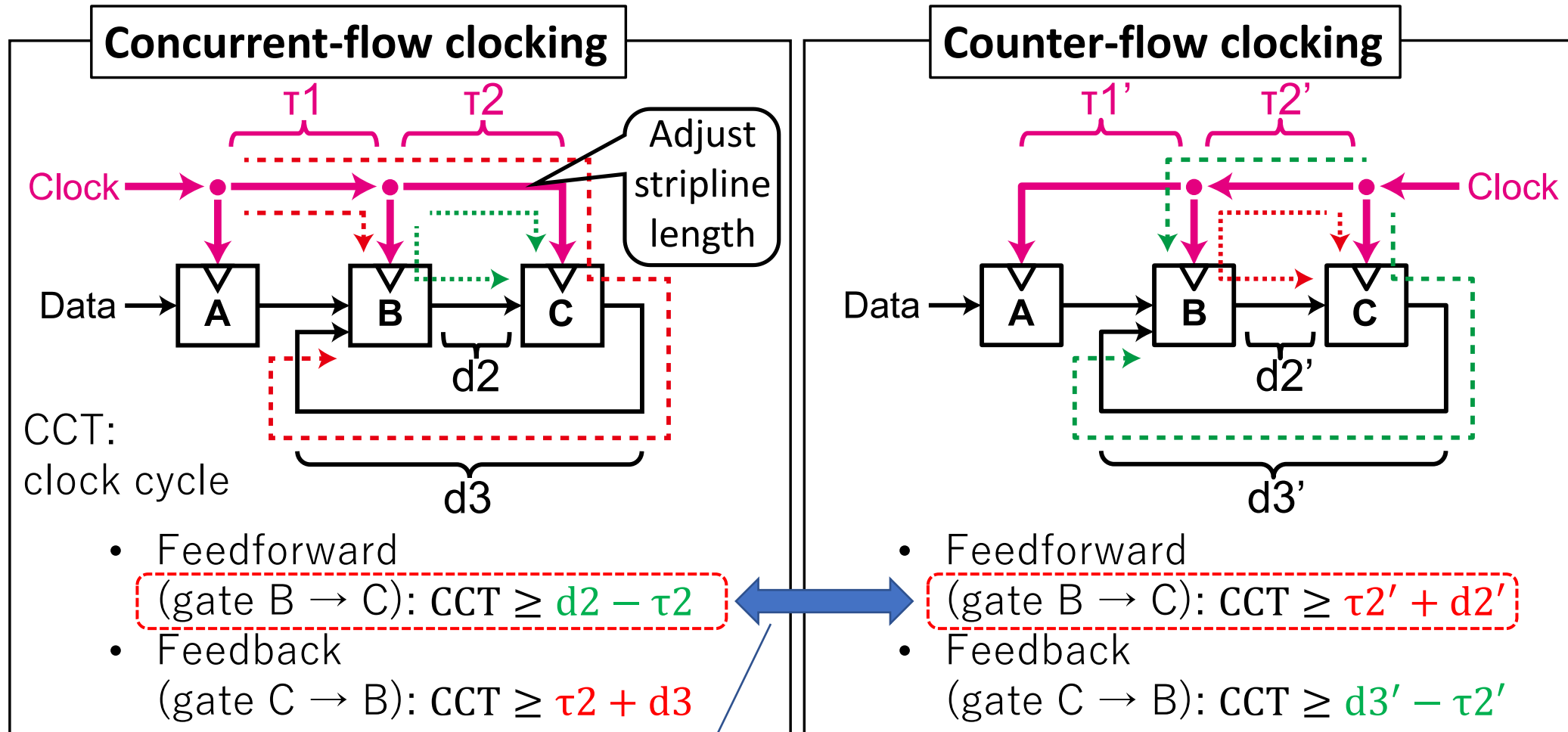
- How to increase frequencies in the circuits containing feedback loops
- How to increase frequencies in the conventional RSFQ circuits
- How to increase frequencies in the more complicated circuits

Datapath : Circuits with Feedback Loops



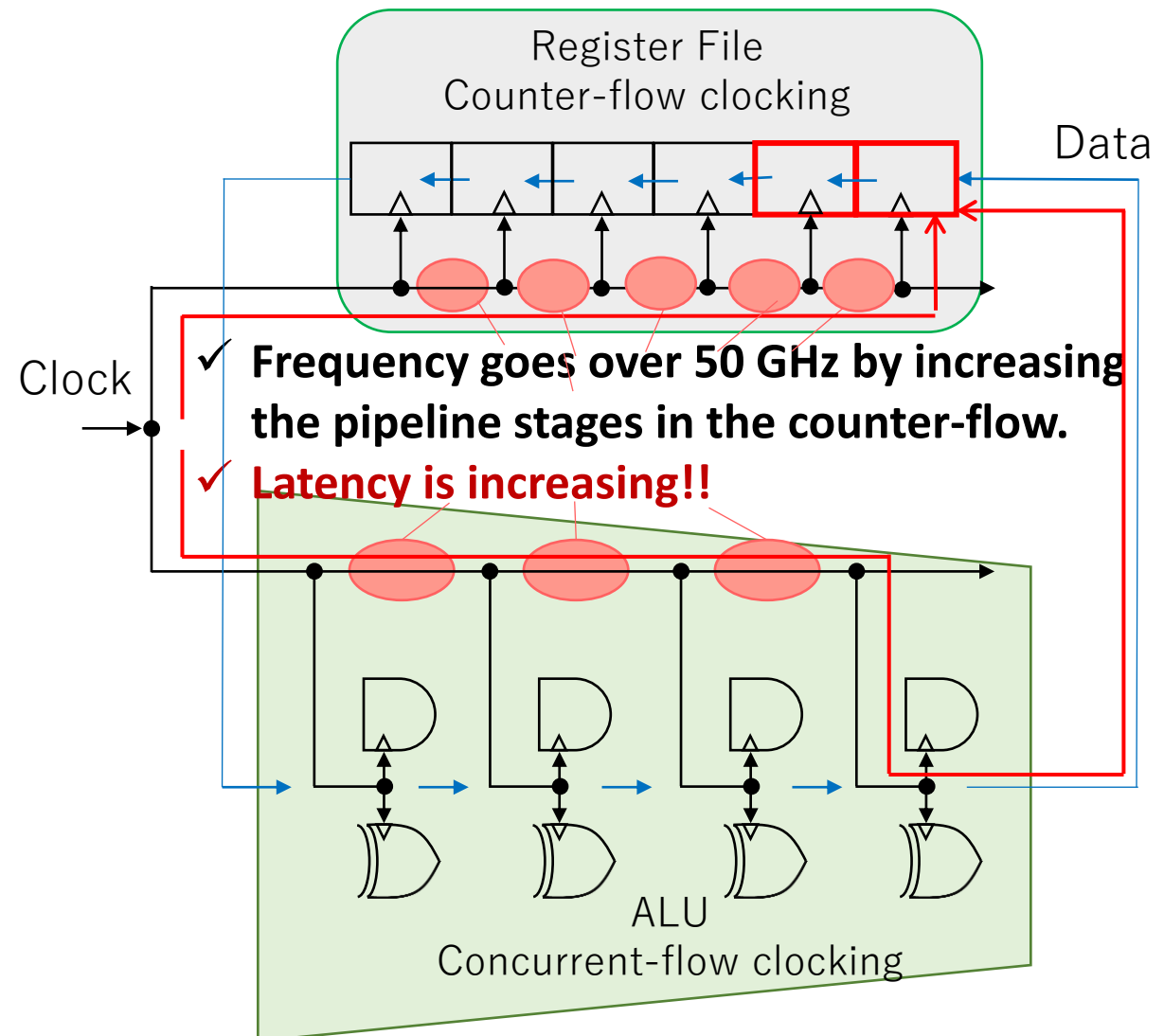
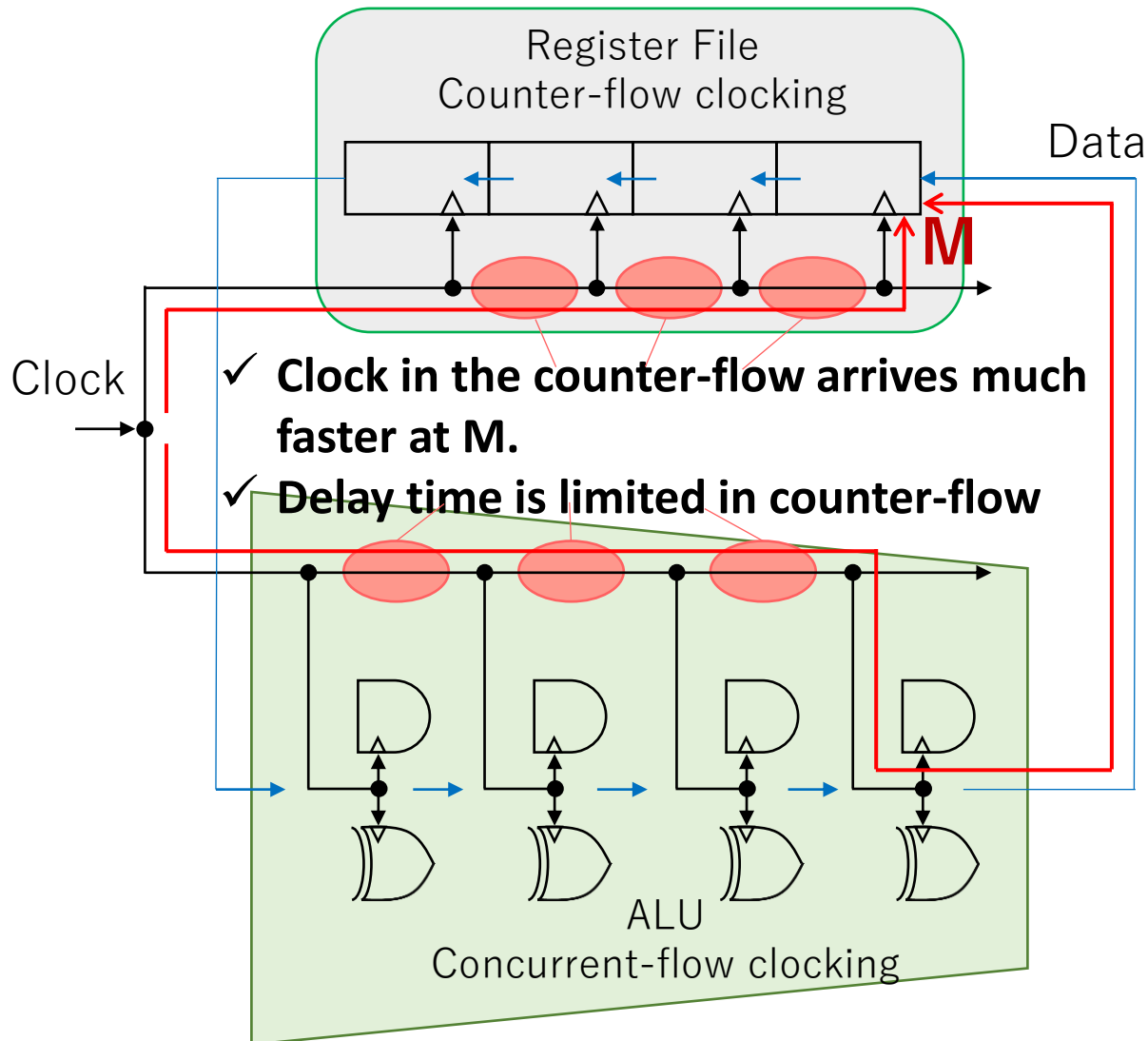
Dual clocking scheme is employed.

Characteristics of Clocking Schemes



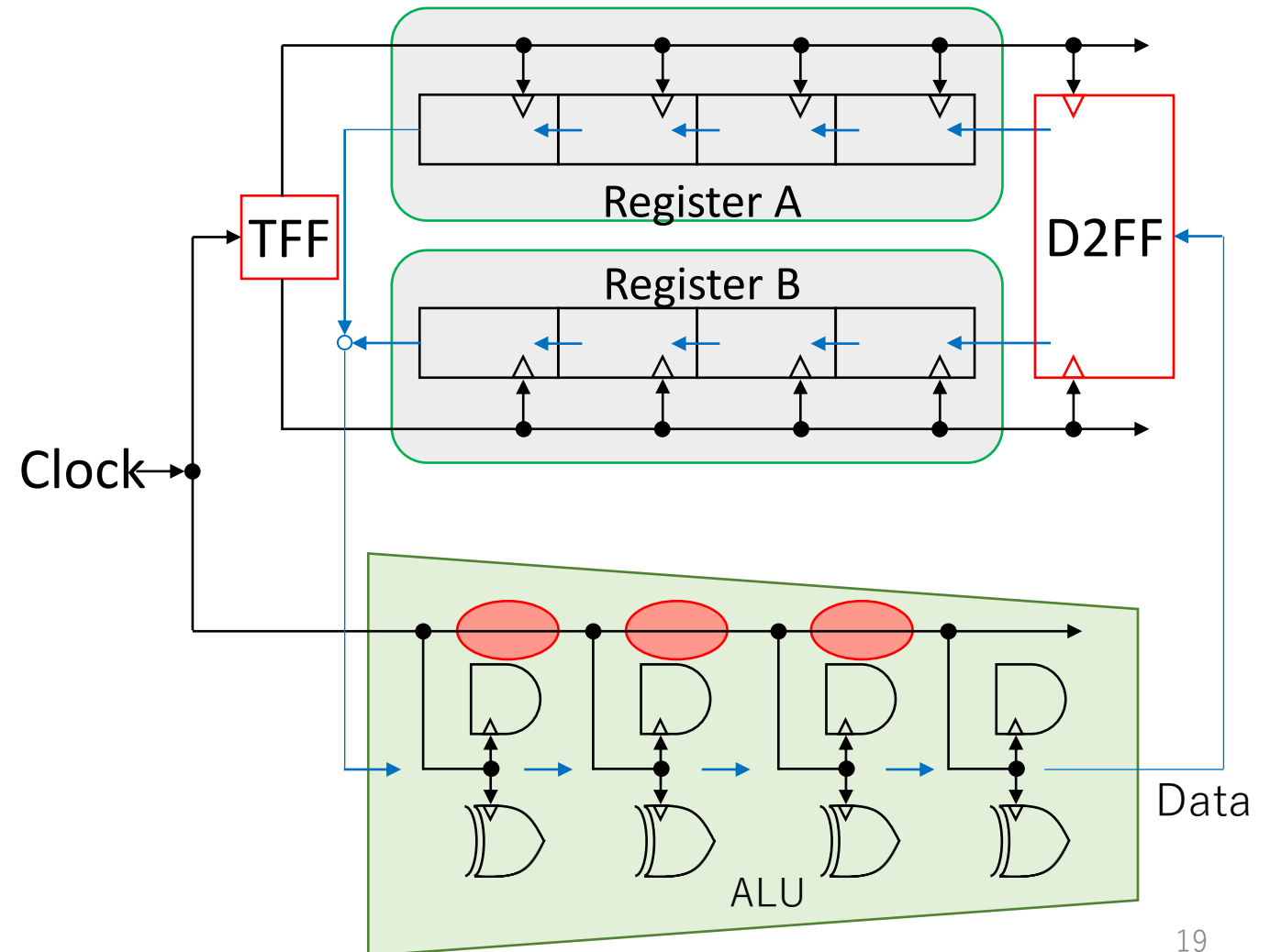
- For raising the frequency in the counter-flow clocking path, τ_2' should be smaller than τ_2 .
- Clock cycle time (CCT) of the counter-flow clocking path is hard to be decreased.

Datapath : First Approach



Interleaved Register File

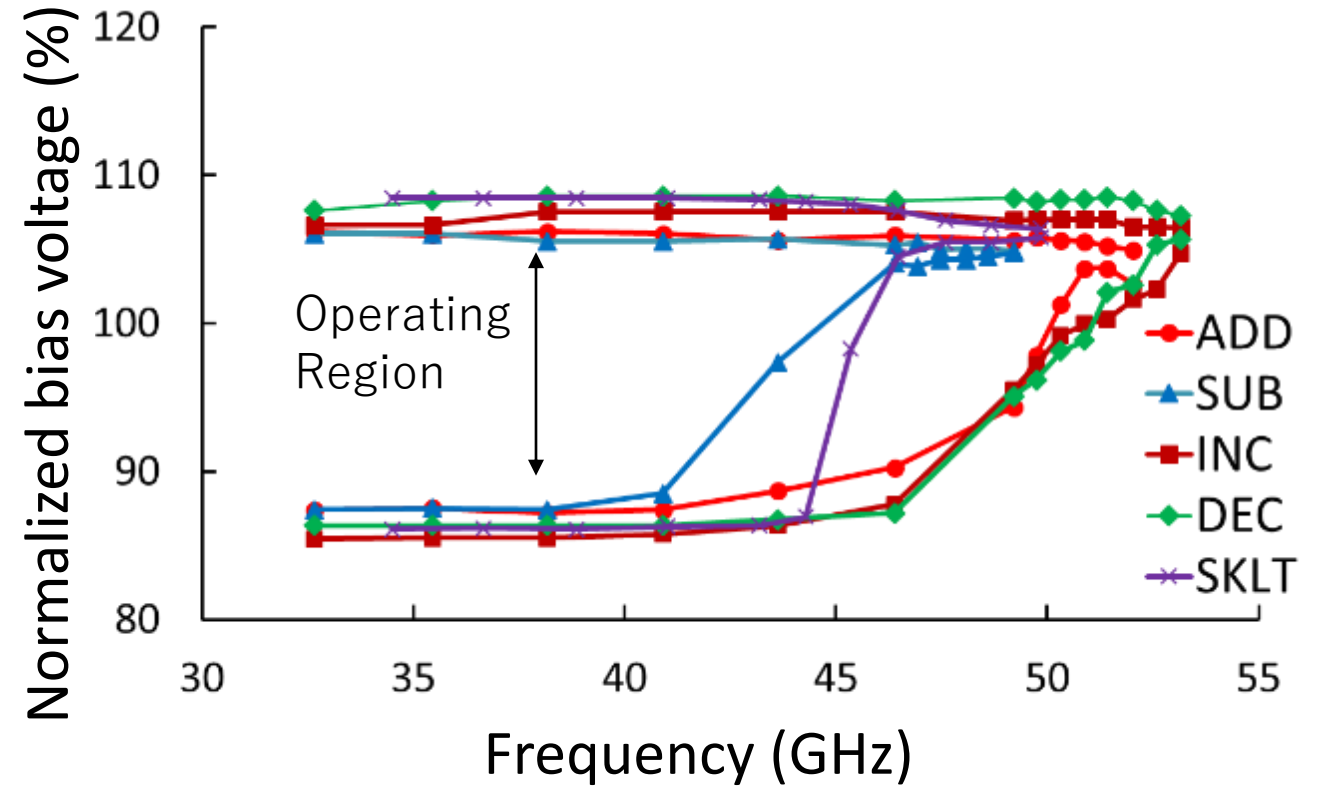
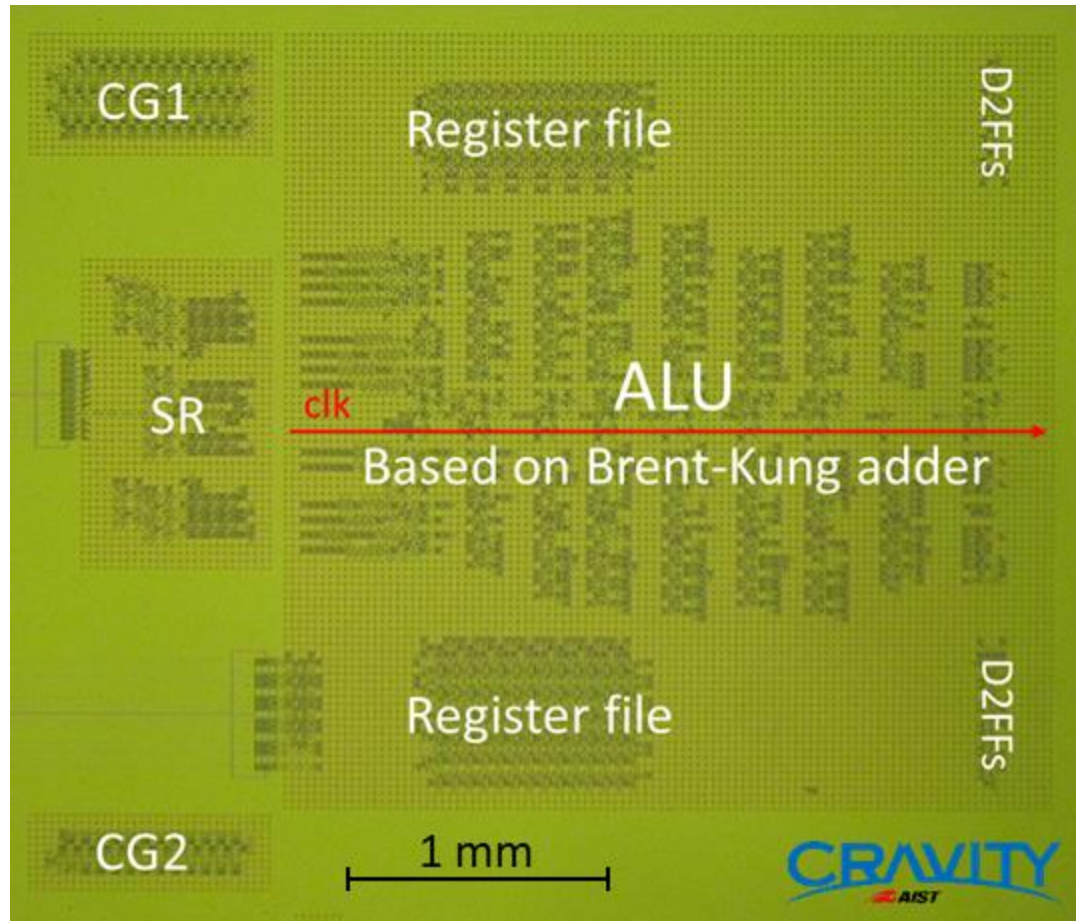
- Clocks are distributed alternately to Register A and B.
- Operating frequency of Registers can be half with keeping the frequency of the ALU.



Estimated Latency

	w/ interleaved registers	w/o interleaved registers
Number of pipeline stages (ALU + D2FF + RF)	24 (9+1+14)	49 (9+0+40)
Latency	20 ps × 10+40 ps × 14 = 760ps	20ps × 49 = 980ps
JJ counts	10333	9785
Area (mm ²)	3.06 × 3.09	3.24 × 2.52

Test Chip: 4-bit ALU and Register Files

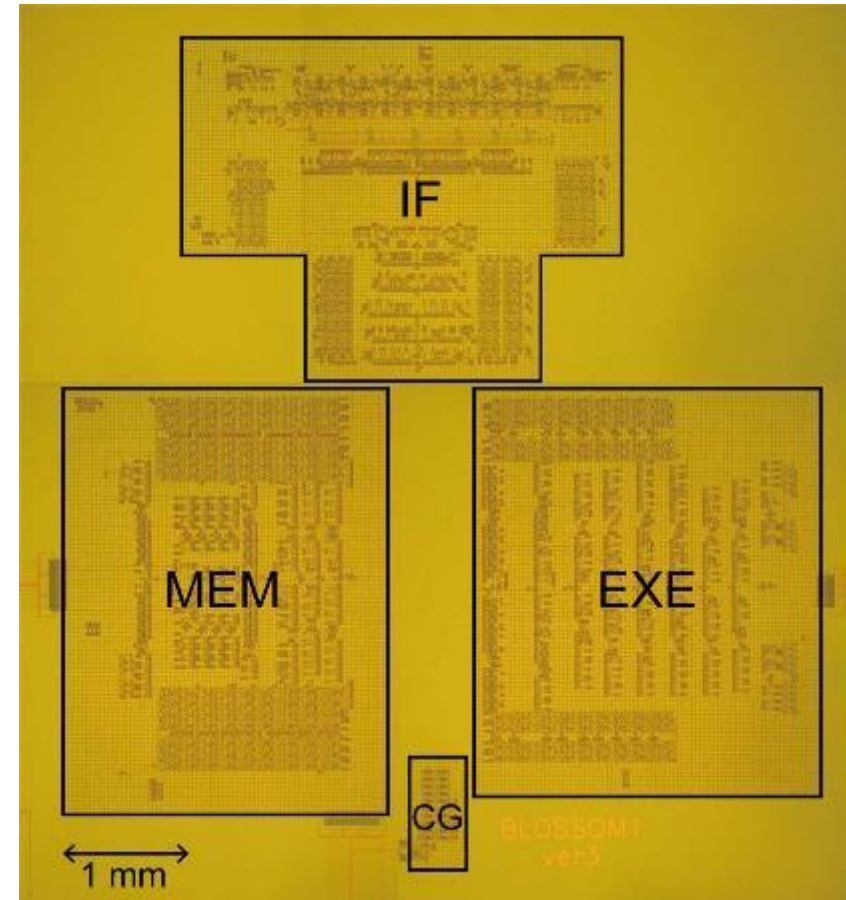
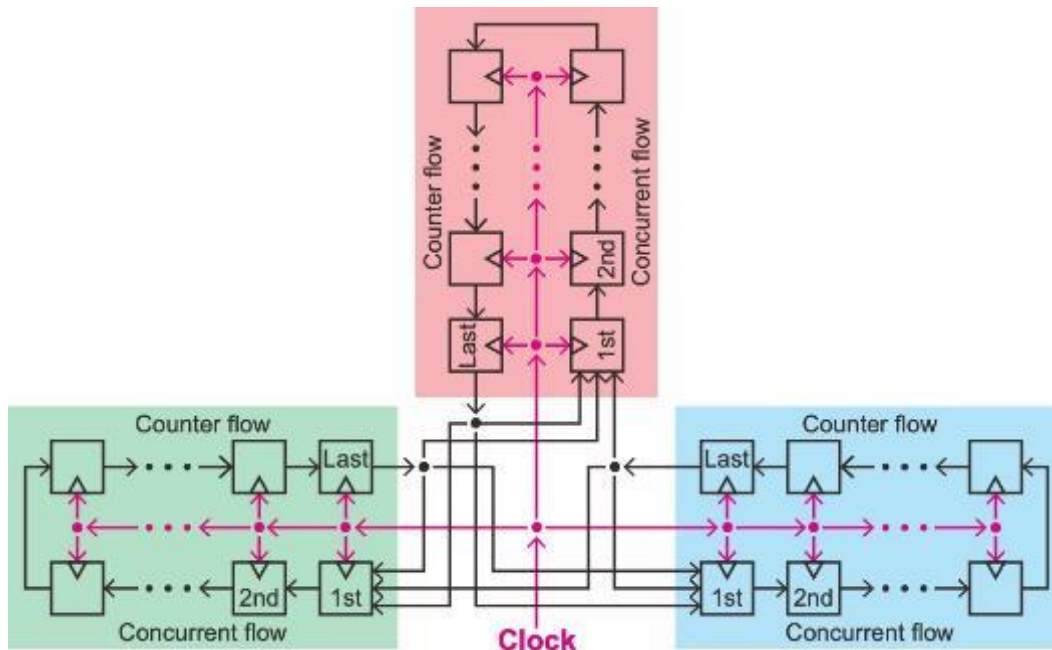


Outline

- Superconductor-based supercomputer in future
- High-speed matrix memory
- **High-speed processor**
 - Way to high-speed operation
 - Low latency in high-speed gate-level-pipelined SFQ datapath
 - **Other SFQ processors**
- Summary

8-bit GLP Microprocessor

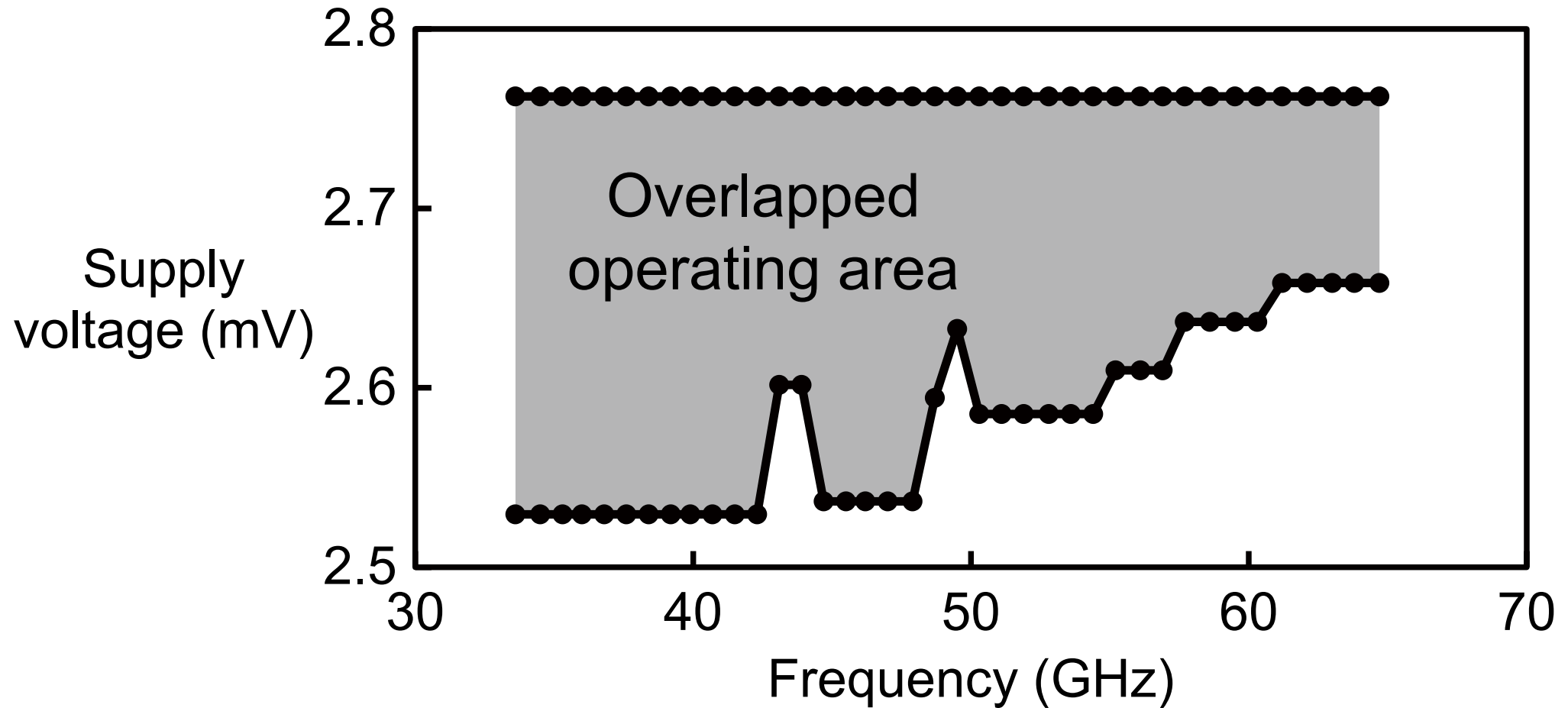
many feedback paths and communications between components



32,712 JJs
 5.8 x 6.0 mm²

To be presented at A-SSCC 2022

Margin of 8-bit GLP Microprocessor

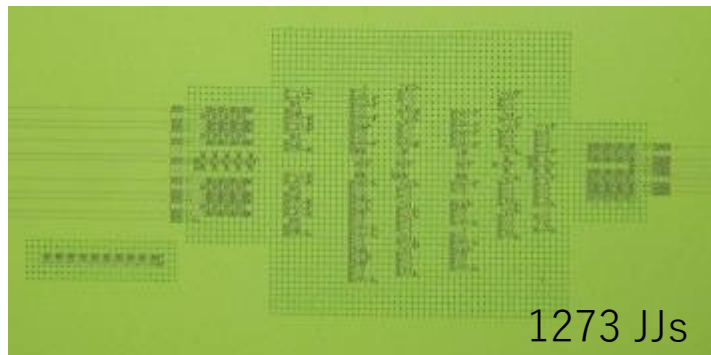
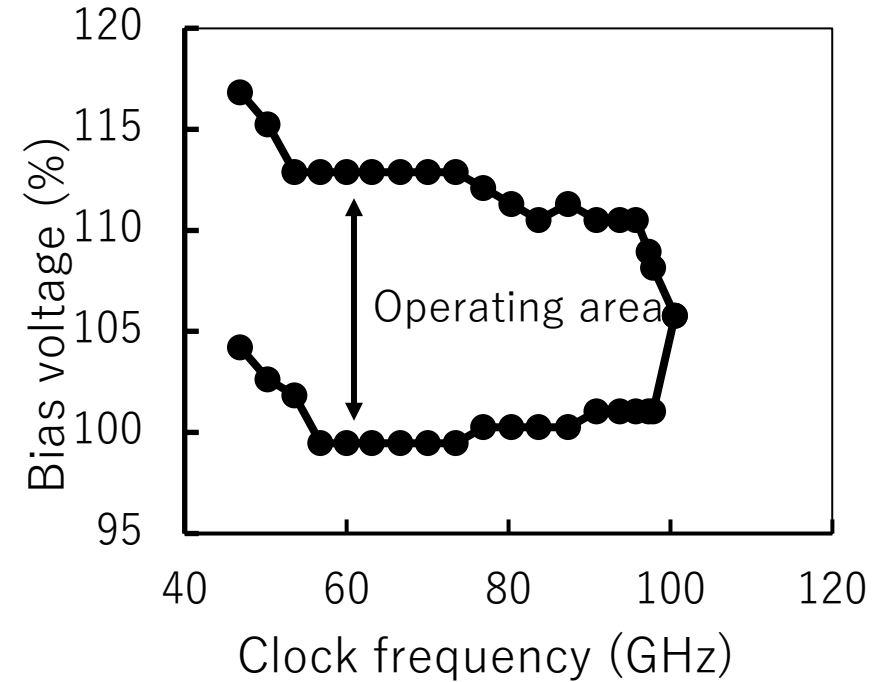
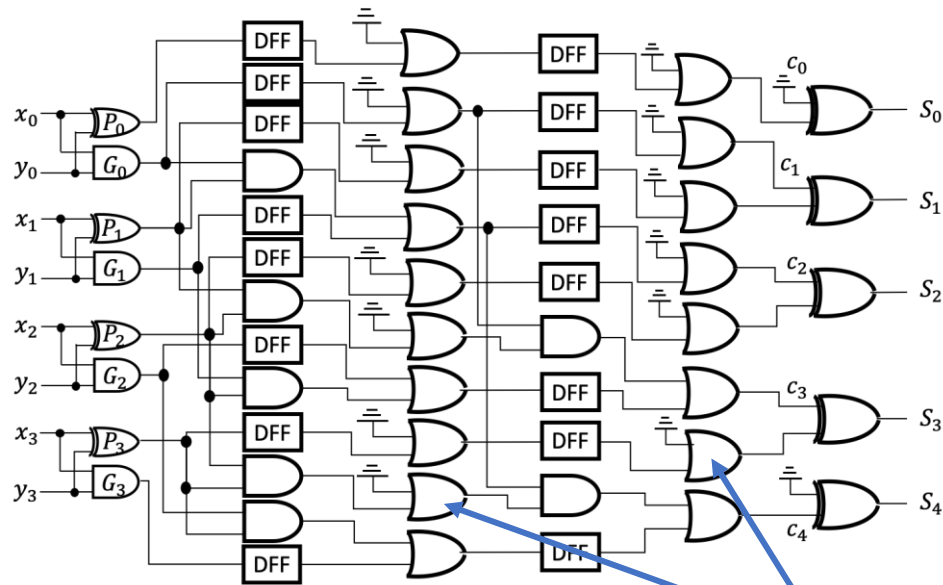


Feedback loops and component communications do not limit frequency.

The detail will be presented by I. Nagaoka at A-SSCC 2022 held November 9th, 2022.

Increase Frequencies in RSFQ Circuits

4-bit Adder



**Dummy gate w/
similar timing
characteristics**

Successfully operated up to 101 GHz

Summary

- Design techniques of the bit-parallel gate-level-pipelined circuits have been refined. The operating frequencies goes over 50 GHz under J_c of 10 kA/cm².
- The keys are as follows
 - Input Data are set at the center of the time window.
 - Gates that have similar timing characteristics are used.
 - For the dual clock scheme employed in the datapath, the interleaved register files are effective for high-speed operation.
- High-speed matrix memories, which have been an issue for more than 50 years, are possibly overcome by using π -junctions.