# Performance and Power Consumption Analysis of Arm Scalable Vector Extension with gem5 Simulator

Tetsuya Odajima, Yuetsu Kodama, Miwako Tsuji, Mitsuhisa Sato

RIKEN R-CCS, Japan

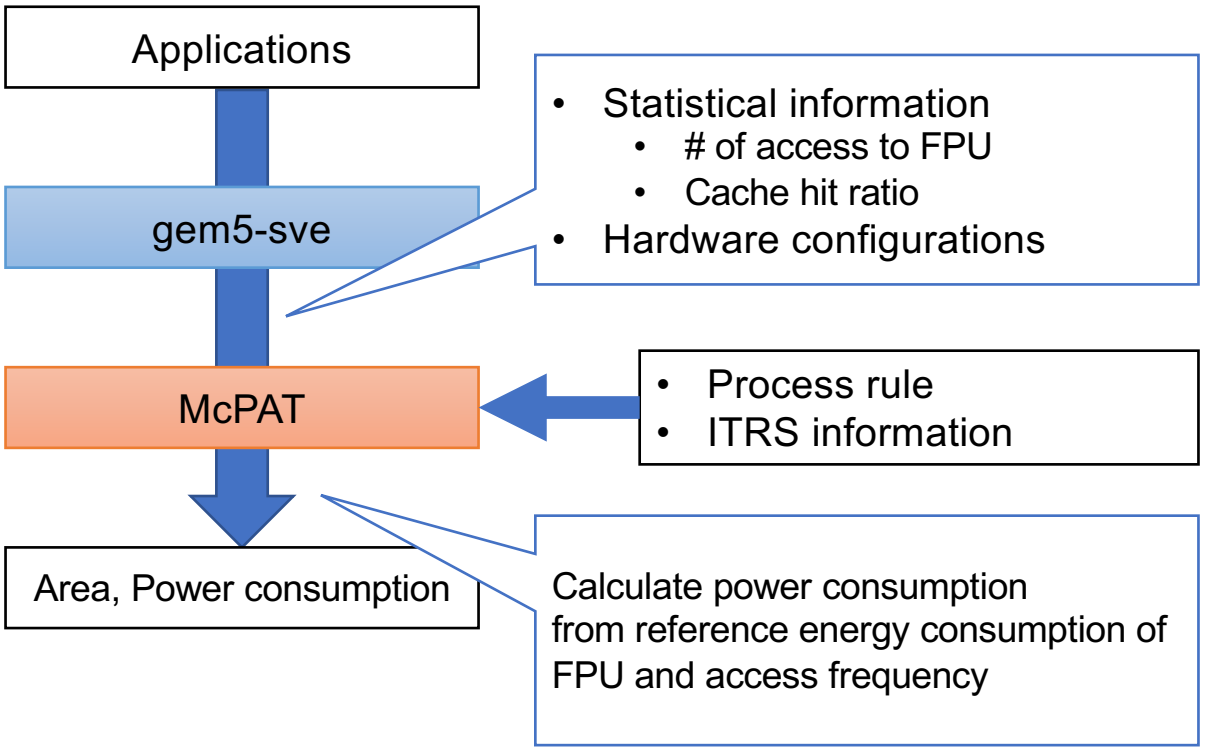
# Background of Our Research

- We have been working on the performance evaluation and application tuning for A64FX processor of Post-K by the simulator
- **Arm Scalable Vector Extension (SVE)**
  - Support from 128-bit to 2048-bit vector length (VL)
  - Vector Length Agnostic (VLA) ; there is no VL information in assembly code

- Working on research of processor architecture by using simulator
  - Simulator can change its various hardware parameters easier
  - We use alternative parameters, not that of the Post-K
- Our simulator environment
  - **gem5-sve**: general-purpose processor simulator, supports SVE
    - We have been implementing some instructions
  - **McPAT**: framework for calculating power consumption
    - Estimate the power consumption from the # of access to FPU, registers, caches, etc.

# Our Simulator Environment

- Flow of estimating area and power consumption



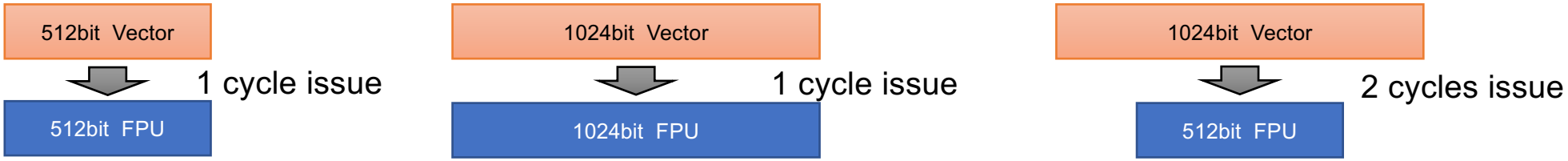
# Purpose of Our Research

- We will evaluate the effect of vector length in the performance and the energy consumption by using simulator

- Comparing the performance and energy consumption among 512-bit and 1024-bit vector length
  - Peak performance of 1024-bit FPU is twice that of 512-bit

  → Area of FPUs and register fires also will be increasing nearly twice
- Various implementations of FPU

Peak performance is half
↕ trade-off
Area of FPU is also half




- We will evaluate the advantage of a longer vector length with multi-cycle FPU

# Evaluation Environment

- Based on ThunderX2 parameter
- We defined instruction latency for SVE referred to NEON

| Hardware parameters | | | |
|---|---|---|---|
| Clock Frequency | 2.1 GHz | # of cores | 1 |
| L1 Dcache, Icache size | 32 kB | L2 cache size | 8 MB |
| Integer pipeline | 1 | Load/Store unit | 1 / 1 |
| Floating pipeline | 2 | Process rule | 32 nm |

| Out-of-Order resource parameters | |
|---|---|
| IQ (Reservation Station) | 60 |
| ROB (Re-order Buffer) | 180 |
| LQ (Load Queue) | 16 |
| SQ (Store Queue) | 16 |
| Physical Vector Register | 144 |

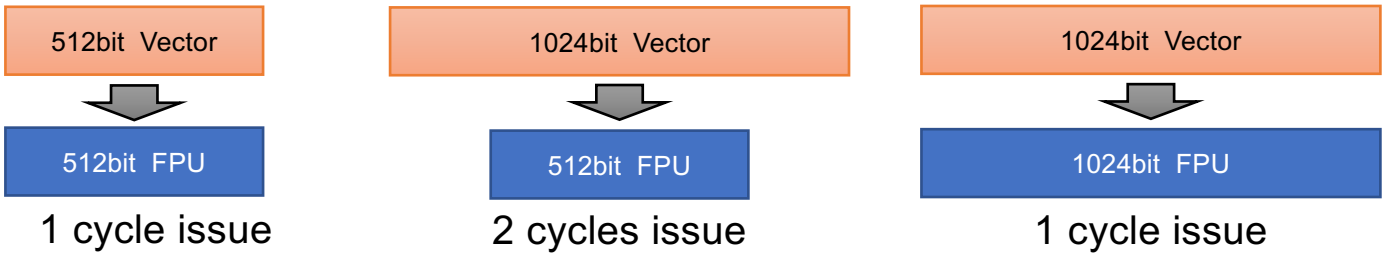
# Evaluation Environment (Cont'd)

- Compiler
  - gcc-arm 8.2, -Ofast -march=armv8-a+sve
- Kernels
  - N-body: latency-bound
    - N=512, TIME_STEP=1
  - Stream Triad: cache/memory-bound
    - N=25600, TIME_STEP=10
- In this evaluation, we don't consider "leak current"

# Evaluation Environment (Cont'd)

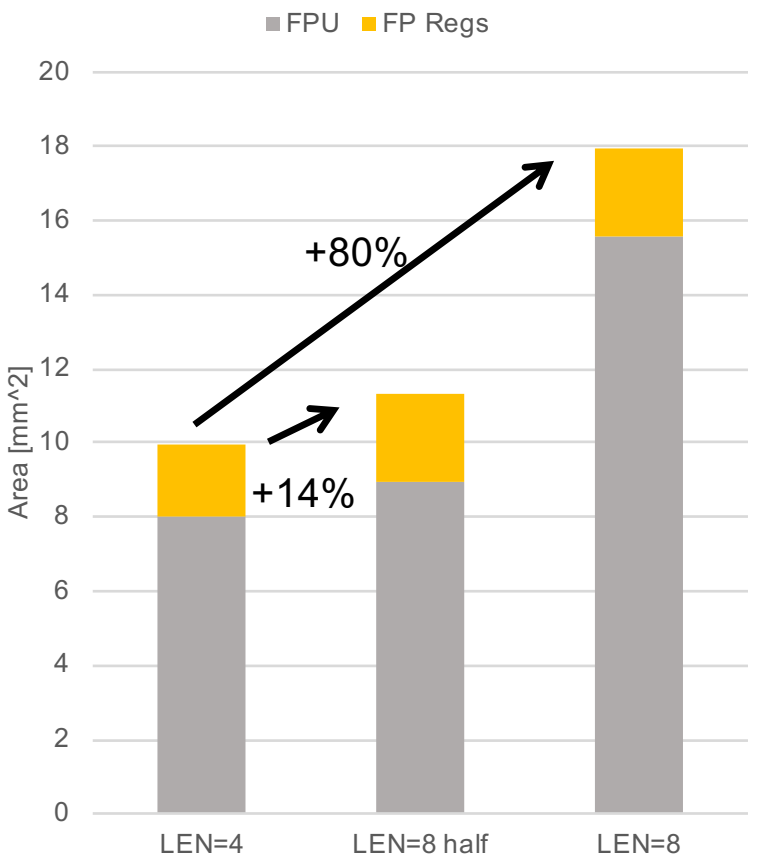
- In gem5, SIMD width of execution unit and register file will be changed by the VL

| | LEN=4 | LEN=8 half | LEN=8 |
|---|---|---|---|
| Vector Length | 512-bit | 1024-bit | → |
| FPU throughput | 512-bit / cycle | 1024-bit / 2cycles | 1024-bit / cycle |
| Peak FPU performance ratio | 1 | 1 | 2 |
| Number of registers | 512-bit x 144 | 1024-bit x 144 | → |

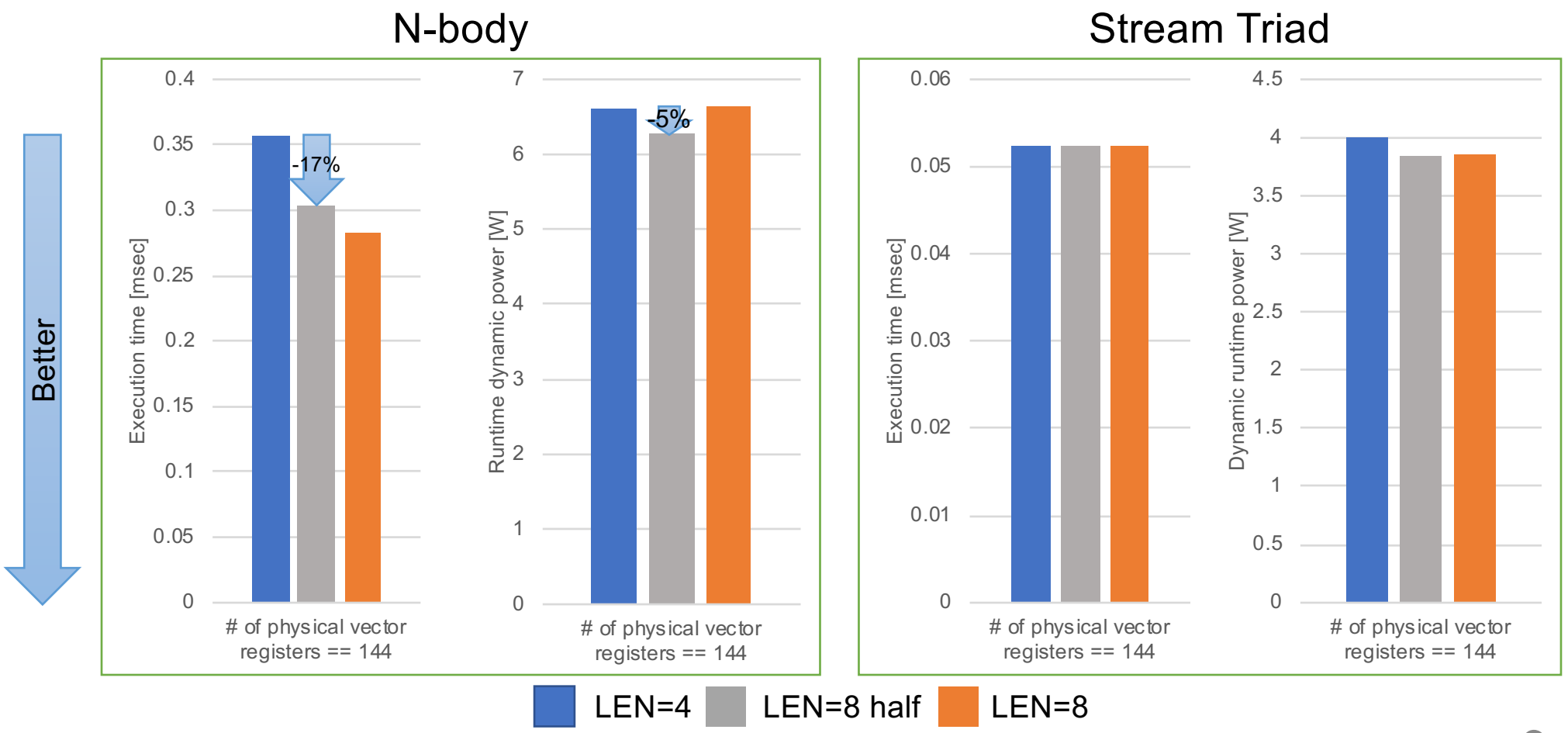


# Hardware Area

- Result of McPAT
- Focus on only FPU & FP Reg. files
- Comparing to LEN=4,
  - area of LEN=8 is +80%
  - area of LEN=8 half is only +14%
- The impact to multi-core is great



# Evaluation | Time & Energy Consumption



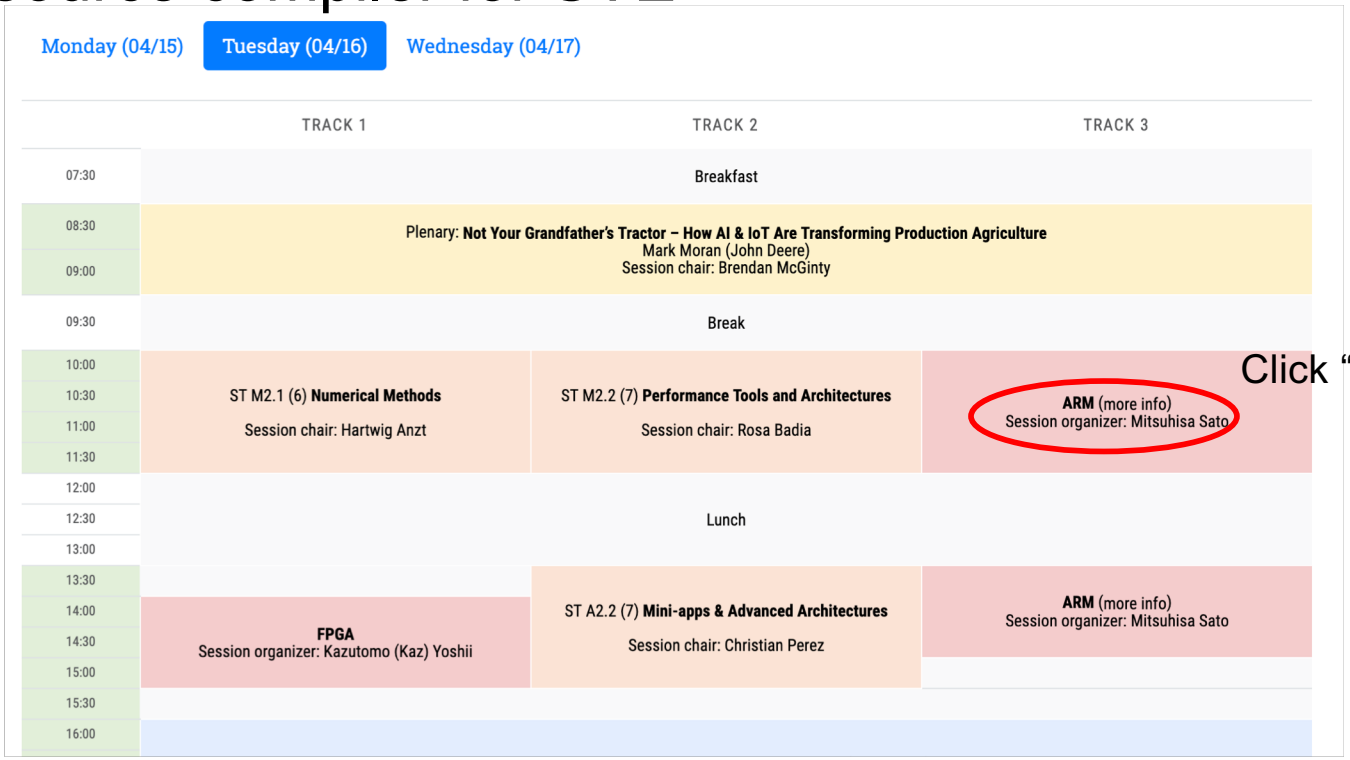
# Conclusion

- By increasing the vector length, performance improvement and low power consumption can be realized in latency-bound kernel
  - N-body : "LEN=8 half" achieves **17% speedup** than LEN=4
    **5% low energy**
- Vector length dose not affect in the performance of memory-bound kernel
- The energy consumption of "LEN=8 half" is almost the same as that of LEN=8

- A longer vector length with multi-cycle vector units (LEN=8 half) is well balanced between the performance and hardware resource

# Download Our Simulator

- We provide Docker image includes our simulator and open source compiler for SVE



Monday (04/15) | **Tuesday (04/16)** | Wednesday (04/17)

| | TRACK 1 | TRACK 2 | TRACK 3 |
|---|---|---|---|
| 07:30 | Breakfast | | |
| 08:30 | Plenary: **Not Your Grandfather's Tractor – How AI & IoT Are Transforming Production Agriculture**<br>Mark Moran (John Deere)<br>Session chair: Brendan McGinty | | |
| 09:00 | | | |
| 09:30 | Break | | |
| 10:00 | ST M2.1 (6) **Numerical Methods**<br>Session chair: Hartwig Anzt | ST M2.2 (7) **Performance Tools and Architectures**<br>Session chair: Rosa Badia | **ARM** (more info)<br>Session organizer: Mitsuhisa Sato |
| 10:30 | | | |
| 11:00 | | | |
| 11:30 | | | |
| 12:00 | Lunch | | |
| 12:30 | | | |
| 13:00 | | | |
| 13:30 | | ST A2.2 (7) **Mini-apps & Advanced Architectures**<br>Session chair: Christian Perez | **ARM** (more info)<br>Session organizer: Mitsuhisa Sato |
| 14:00 | **FPGA**<br>Session organizer: Kazutomo (Kaz) Yoshii | | |
| 14:30 | | | |
| 15:00 | | | |
| 15:30 | | | |
| 16:00 | | | |

Click "more info"