Arm Research Summit 2019

FUJITSU

shaping tomorrow with you

# Performance Evaluation of SVE Enabled Arm Processor A64FX using Variable Vector Length

Shinji Sumimoto
Fujitsu Limited

# Overview

FUJITSU

- Background
- A64FX Overview
- Application Characteristics:
  Compute Intensive vs. Memory Intensive
- Preliminary Performance Evaluation

# Background

FUJITSU

- A64FX is the First SVE enabled Arm Processor in the world.
  - SVE realizes single binary for multiple vector length environment in order to support application binary portability.
  - A64FX supports 512, 384, 256, 128 bit vector length.
- HPC applications have several characteristics such as compute intensive and/or memory intensive.
  - SVE enabled processor can control execution vector length at runtime.
  - A64FX has a memory bandwidth controlling feature at runtime.
  - No re-compilation is needed for the above executions.
- Therefore, we have evaluated several application benchmarks in order to clarify application characteristics

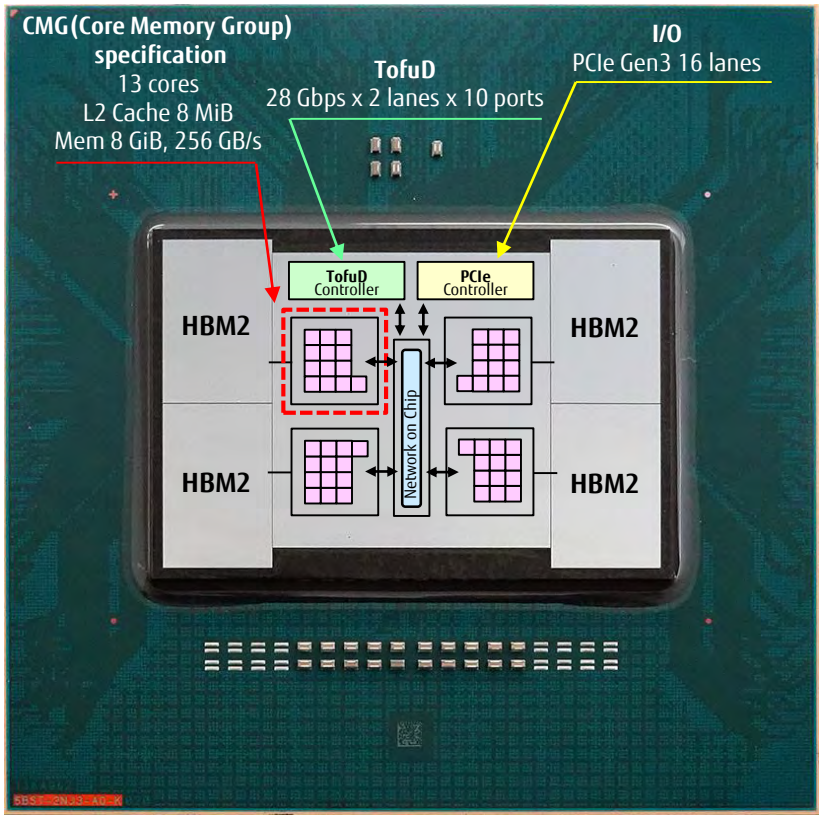
FUJITSU

# A64FX: High Performance Arm CPU

- Inheriting Fujitsu HPC CPU technologies with commodity standard ISA

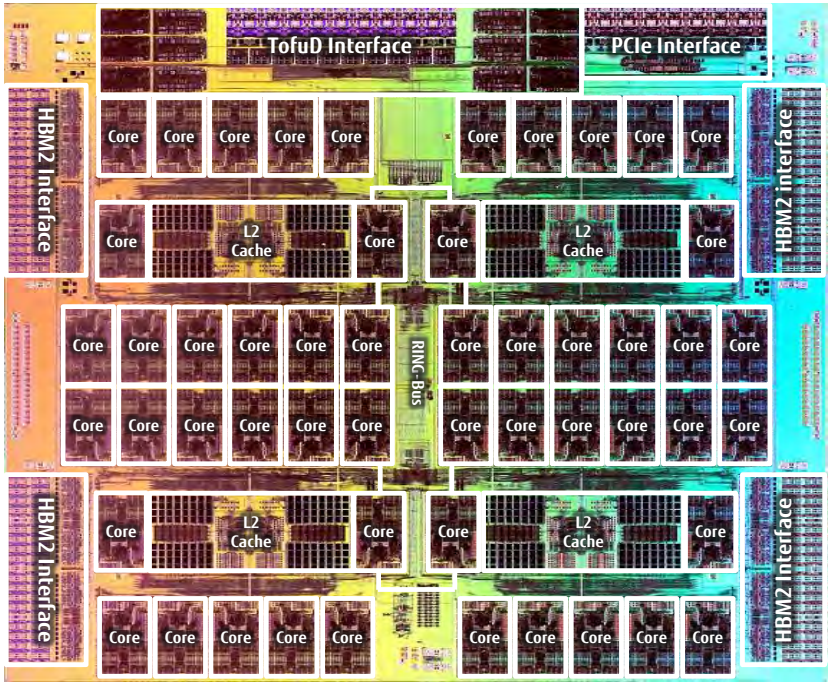
# High Performance Arm CPU “A64FX”

FUJITSU




## Architecture features

| | |
|---|---|
| ISA | Armv8.2-A (AArch64 only) SVE (Scalable Vector Extension) |
| SIMD width | 512-bit |
| Precision | FP64/32/16, INT64/32/16/8 |
| # of cores | 48 computing cores + 4 assistant cores (4 CMGs) |
| Memory | HBM2: Peak B/W 1024 GB/s |
| Interconnect | TofuD: 28 Gbps x 2 lanes x 10 ports |

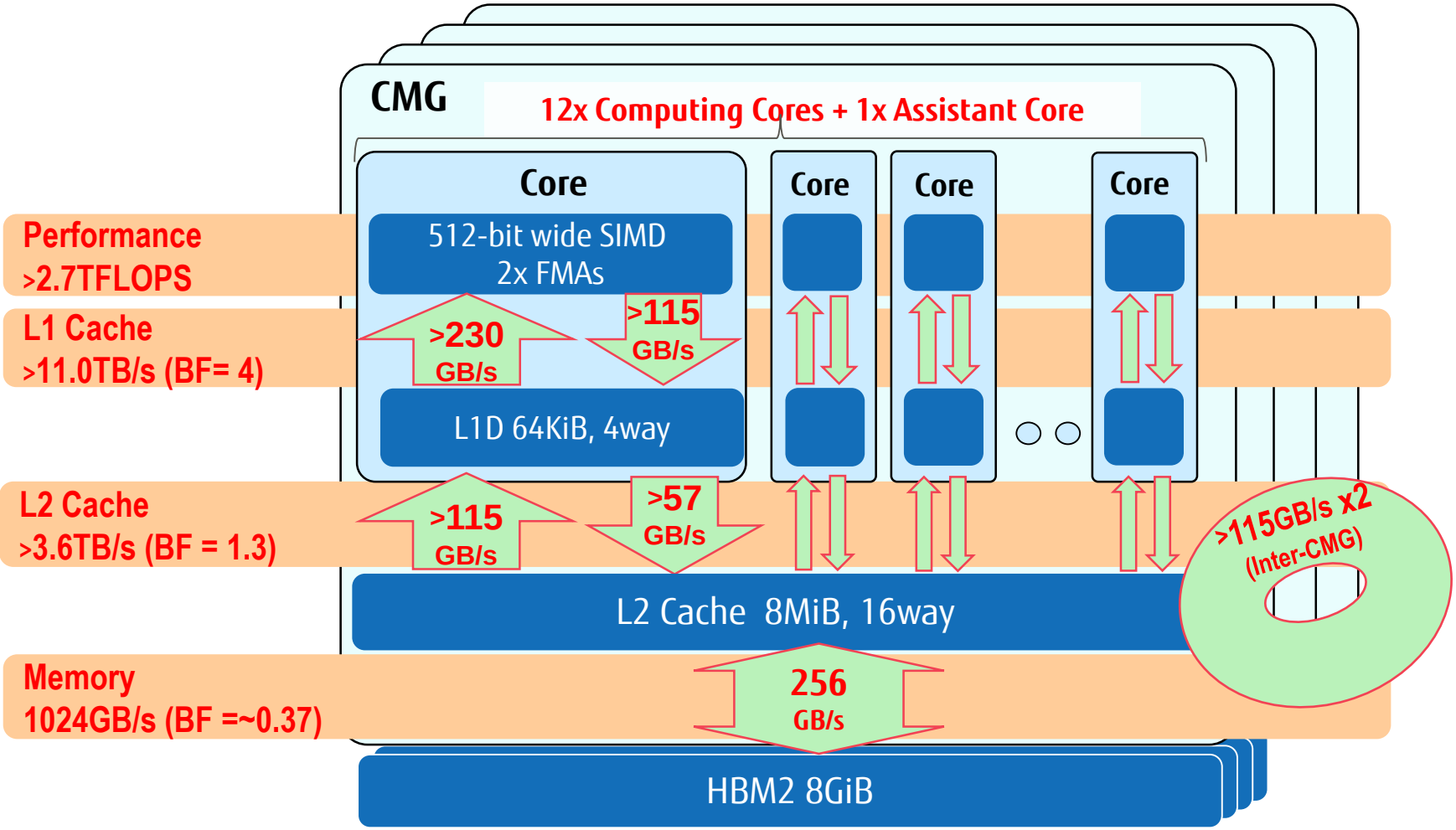


# A64FX Memory System

FUJITSU

## Extremely high bandwidth

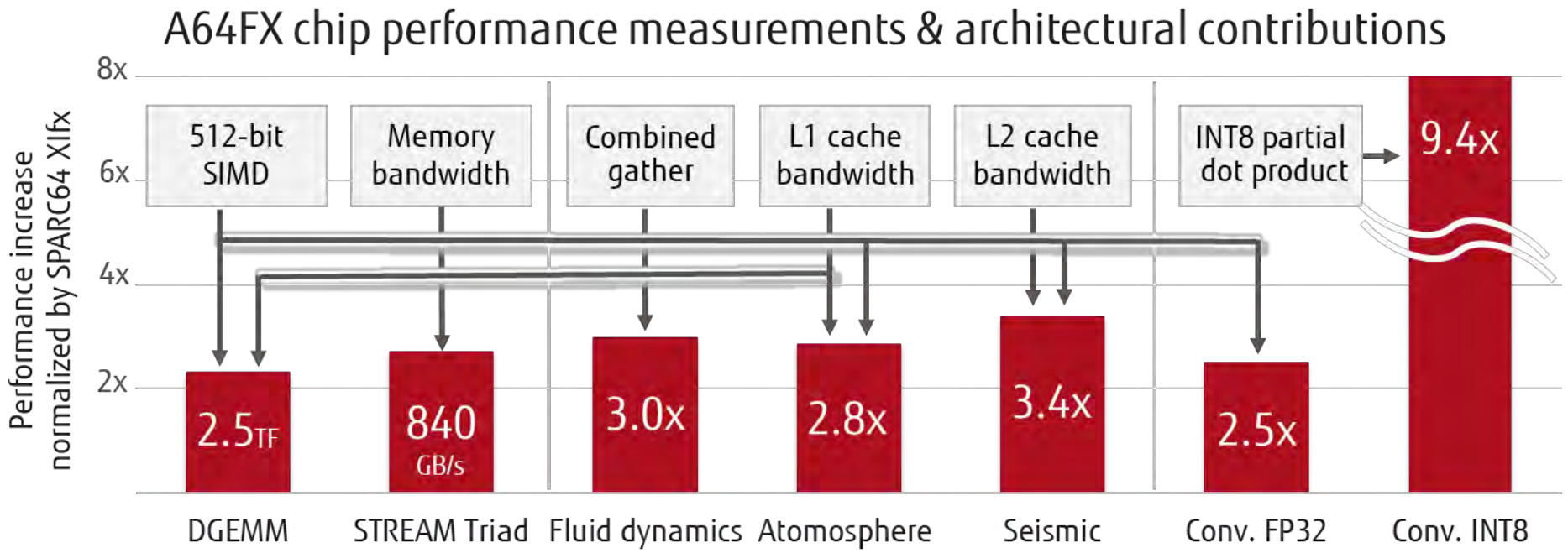
- Asynchronous Processing in cores, caches and memory controllers
- Maximizing the capability of each layer's bandwidth



# A64FX CPU Performance Evaluation



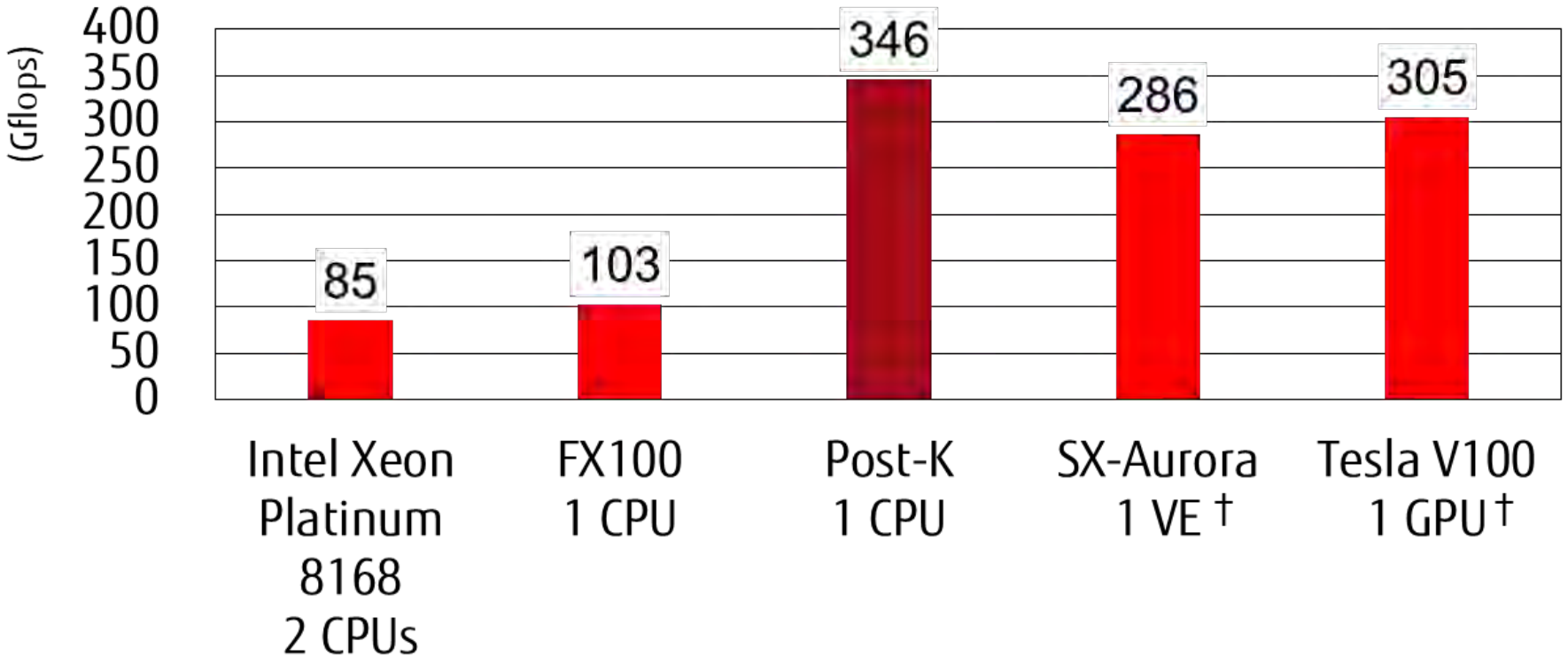
- Over 2.5x faster in HPC & AI benchmarks than SPARC64 XIfx



# A64FX Performance Comparison(1/2)



## Himeno Benchmark (Fortran90)

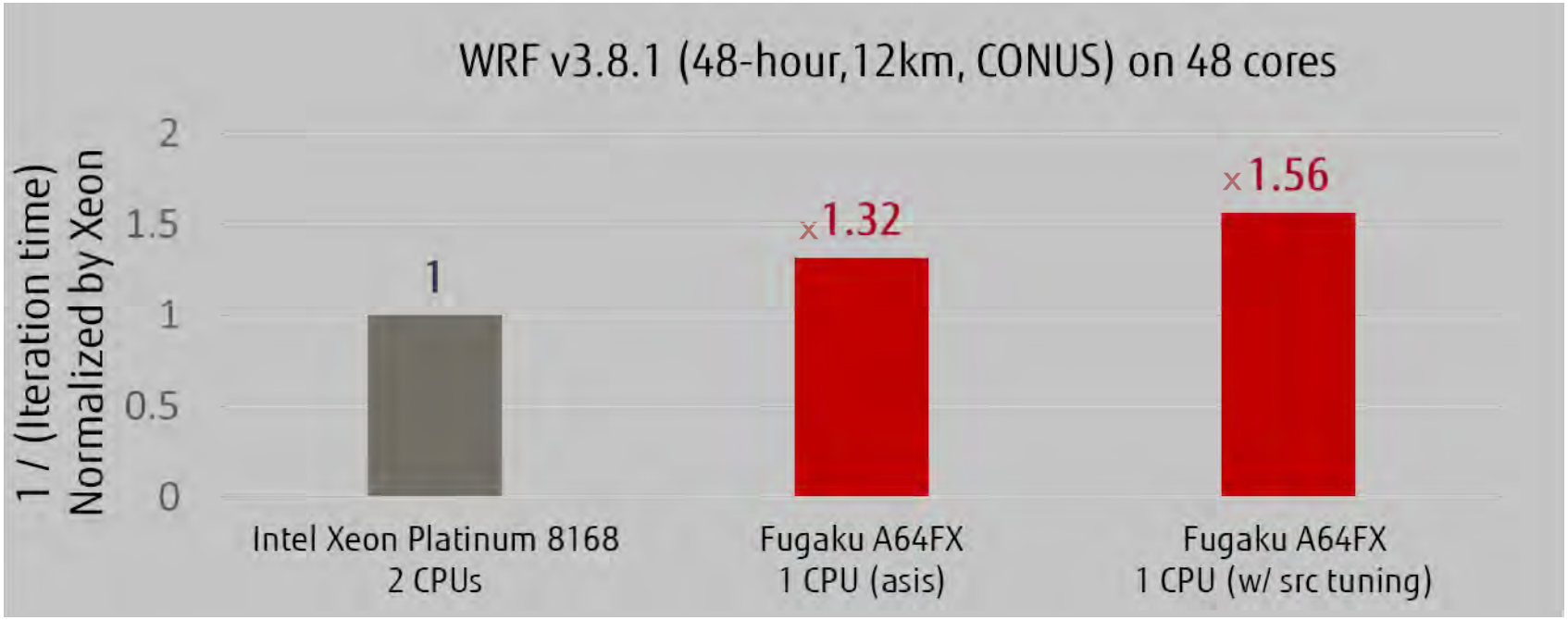



† "Performance evaluation of a vector supercomputer SX-aurora TSUBASA", SC18, https://dl.acm.org/citation.cfm?id=3291728

# A64FX Performance Comparison(2/2)

## WRF: Weather Research and Forecasting model

- Vectorizing loops including IF-constructs is key optimization
- Source code tuning using directives promotes compiler optimizations



FUJITSU

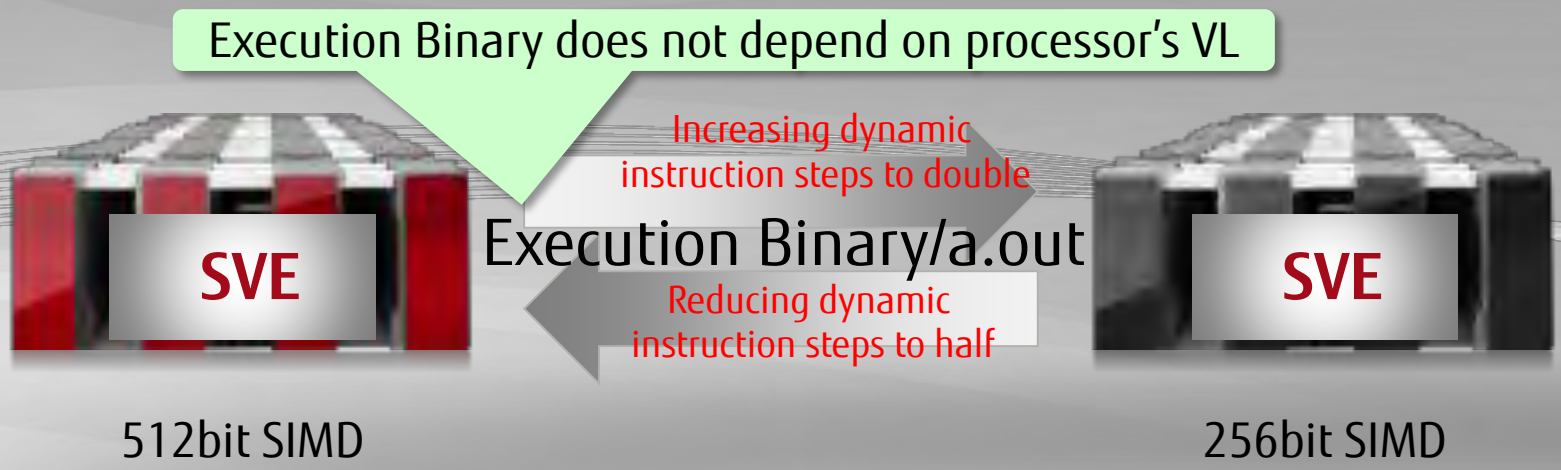
# Application Characteristics: Compute Intensive vs. Memory Intensive

# SVE: Scalable Vector Length

FUJITSU

- ISA does not fix Vector Length
  - SVE supports VL from 128 to 2048 bit with multiples of 128 bit
  - VL is set by processor before executing a binary dynamically
- Single execution binary can be executed on processors with multiple VLs
  - Vector-Length Agnostic(VLA) programing enables ABI Compatibility

**Execution Binary Portability**



# Application Characteristics: How to investigate?

FUJITSU

- Application Characteristics
  - Computing Intensive
  - Memory Bandwidth Intensive
- How to investigate?
  - Using performance profiling tools: Arm Allinea Studio
  - Re-compiling with different compiler option and running
- A64FX can help to evaluate the characteristics easily
  - Compute Intensive Analysis: Changing Vector Length
  - Memory Intensive Analysis:  Changing Memory Access Gap

FUJITSU

# Preliminary Performance Evaluation

# Benchmark Applications, Evaluation Environment and Evaluation Pameters

FUJITSU

- Benchmark Applications
  - STREAM(TRIAD)
  - DGEMM
  - Himeno Benchmarks
  - NAS Parallel Benchmarks OMP Class C (EP, CG, LU, FE, IS, MG, BT, SP)
- Evaluation Environment
  - Hardware: Fugaku-Prototype System with A64FX (single node)
  - Compiler:  Fujitsu Compiler (development version)
- Evaluation Parameters
  - SVE Vector Length:  128 - 512
    - Using a command with prctl(PR_SVE *)
  - Memory Bandwidth: 100%-20%
    - Using an option of submission job

| Application | VL 128 | VL256 | VL512 |
|---|---|---|---|
| HBM100% | | | |
| HBM 80% | | | |
| HBM 60% | | | |
| HBM 40% | | | |
| HBM 20% | | | |

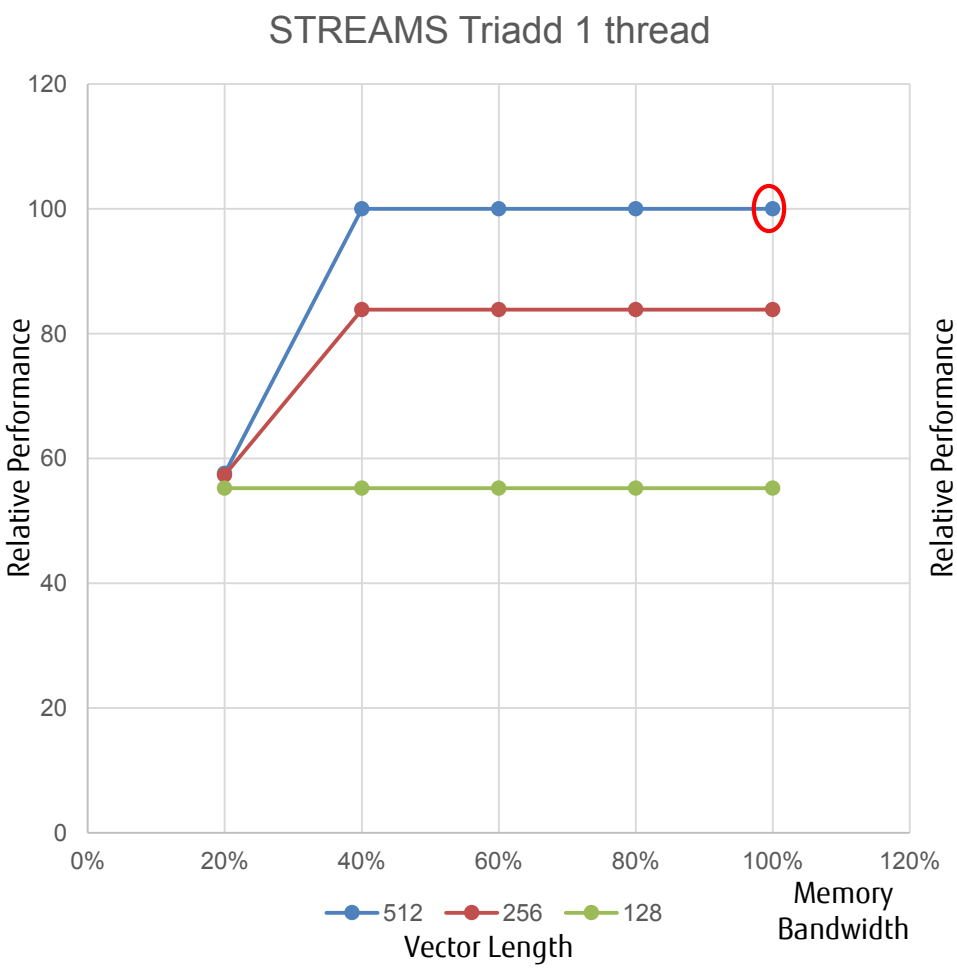
Measured Parameter Combination for each Application

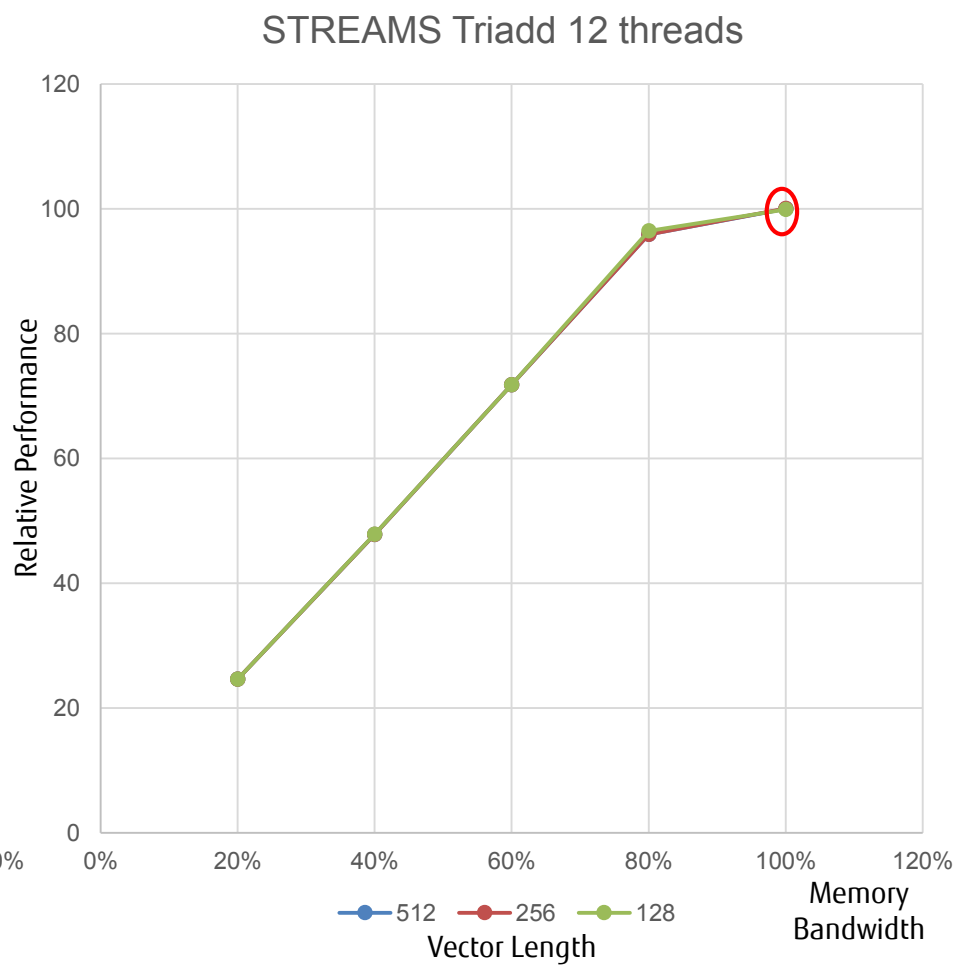
* https://lwn.net/Articles/717804/

# STREAM TRIAD: 1 Thread vs. 12 Threads(1CMG)

FUJITSU

- Relative Performance VL512-HBM100% = 100

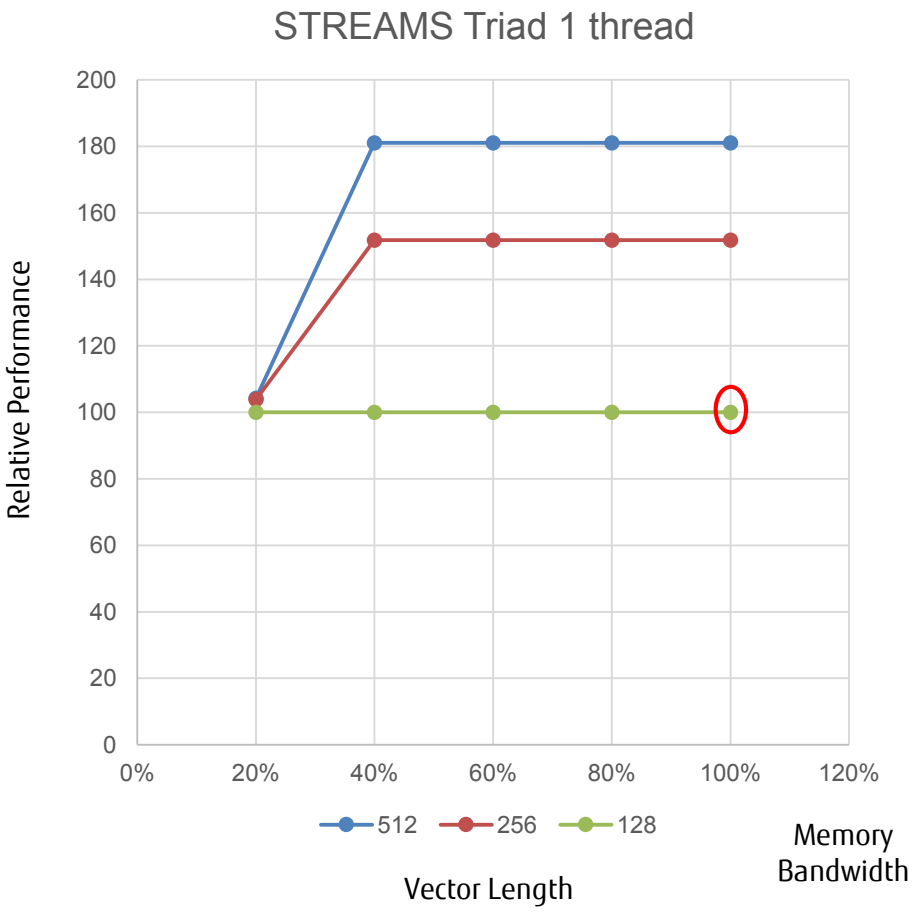


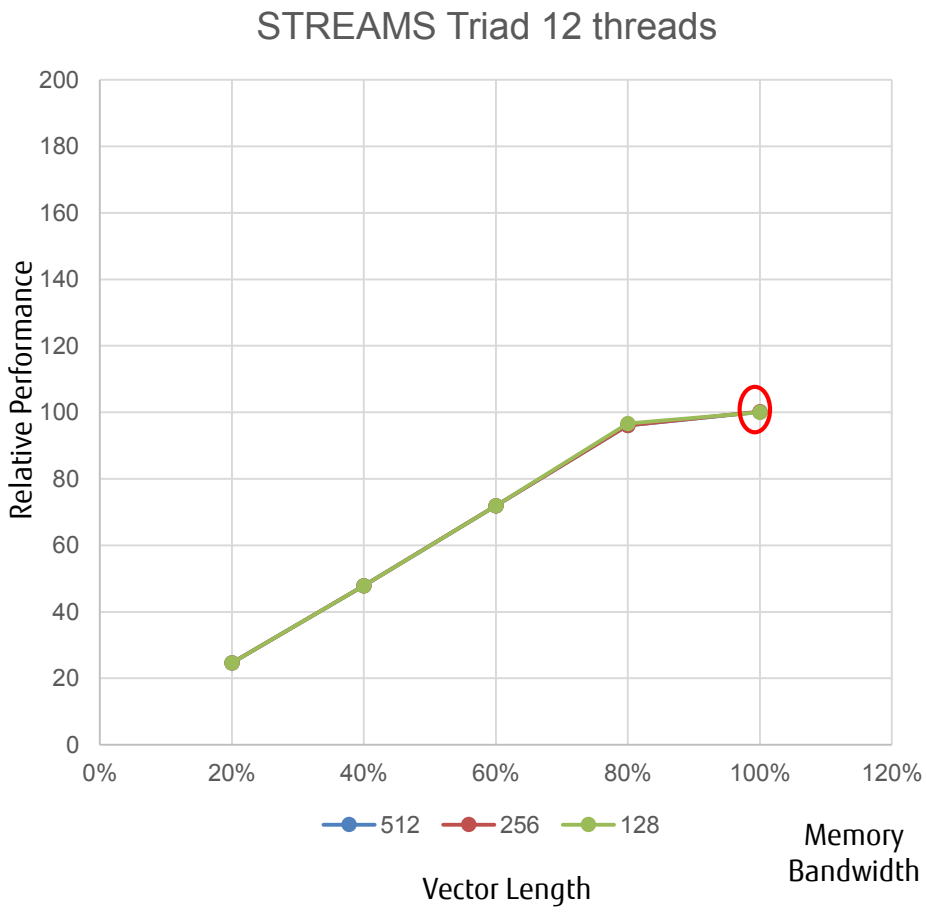



# STREAM TRIAD: 1 Thread vs. 12 Threads-SIMD Effects

FUJITSU

- Relative Performance VL128-HBM 100% = 100

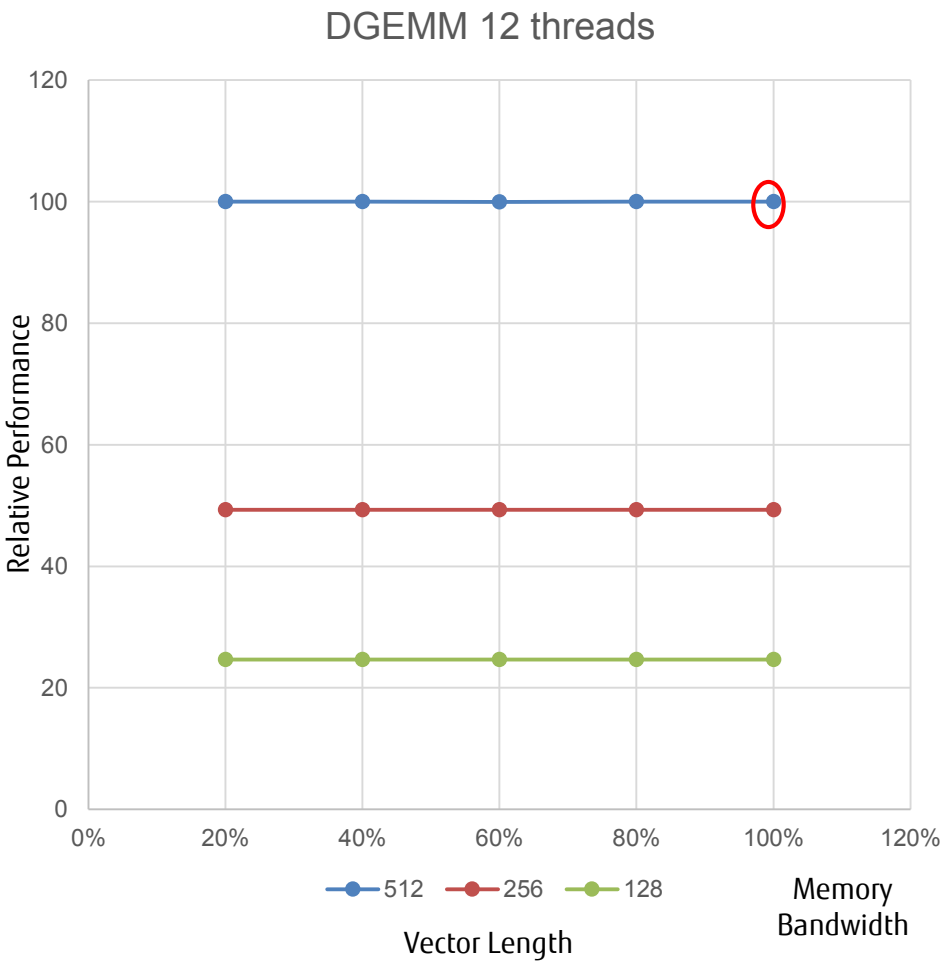


# DGEMM: 12 Threads(1CMG)

FUJITSU

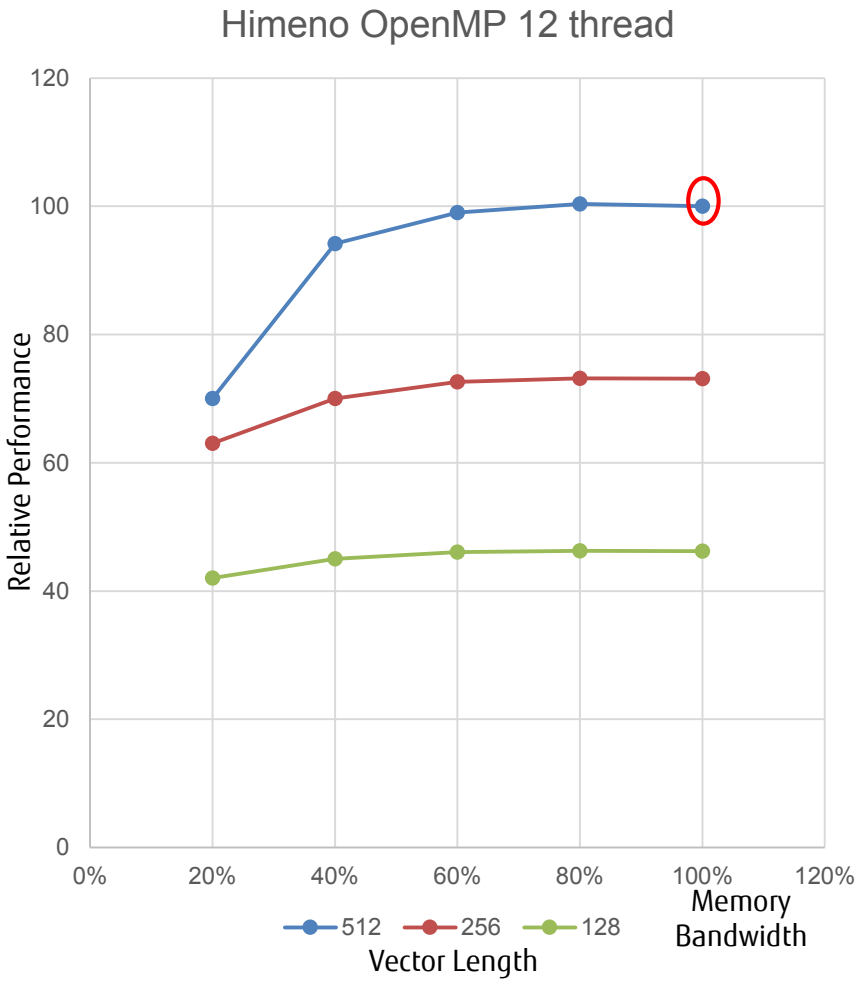
- Relative Performance VL512-HBM100% = 100

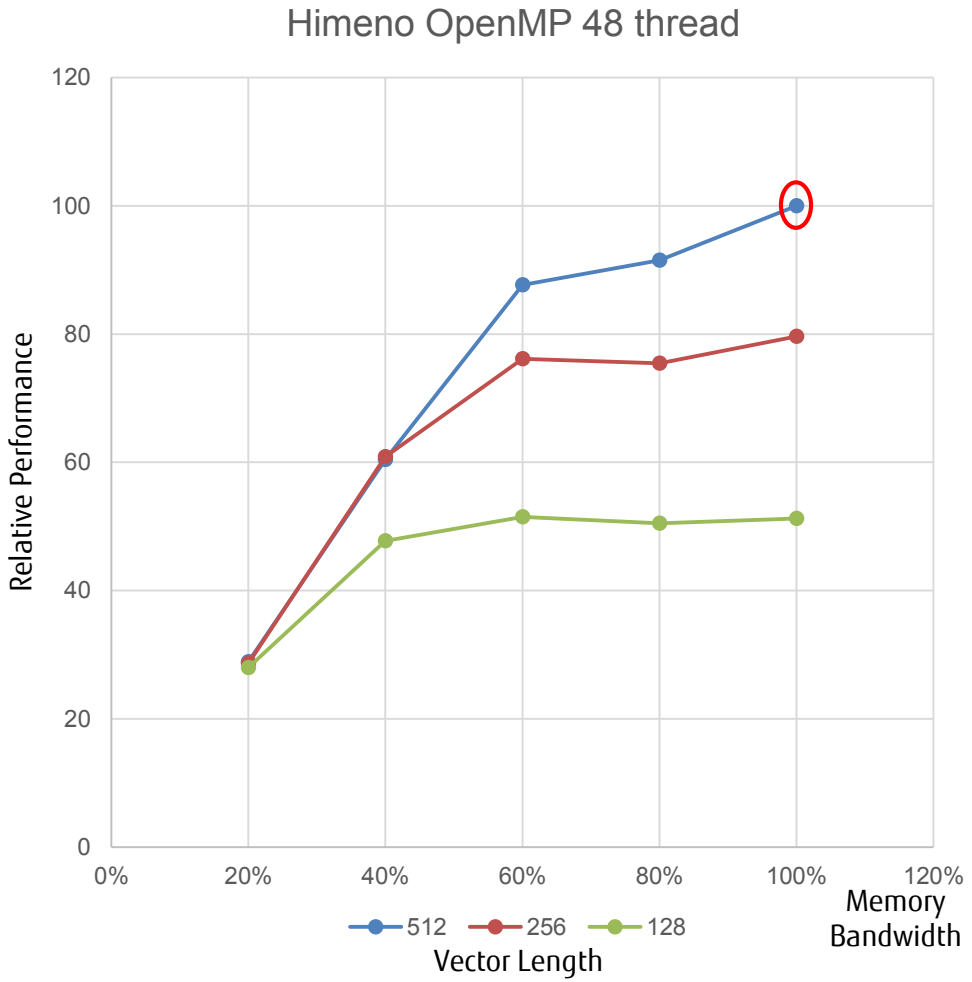


# Himeno Benchmark: 1 CMG vs. 4 CMG

FUJITSU

- Relative Performance VL512-HBM100% = 100

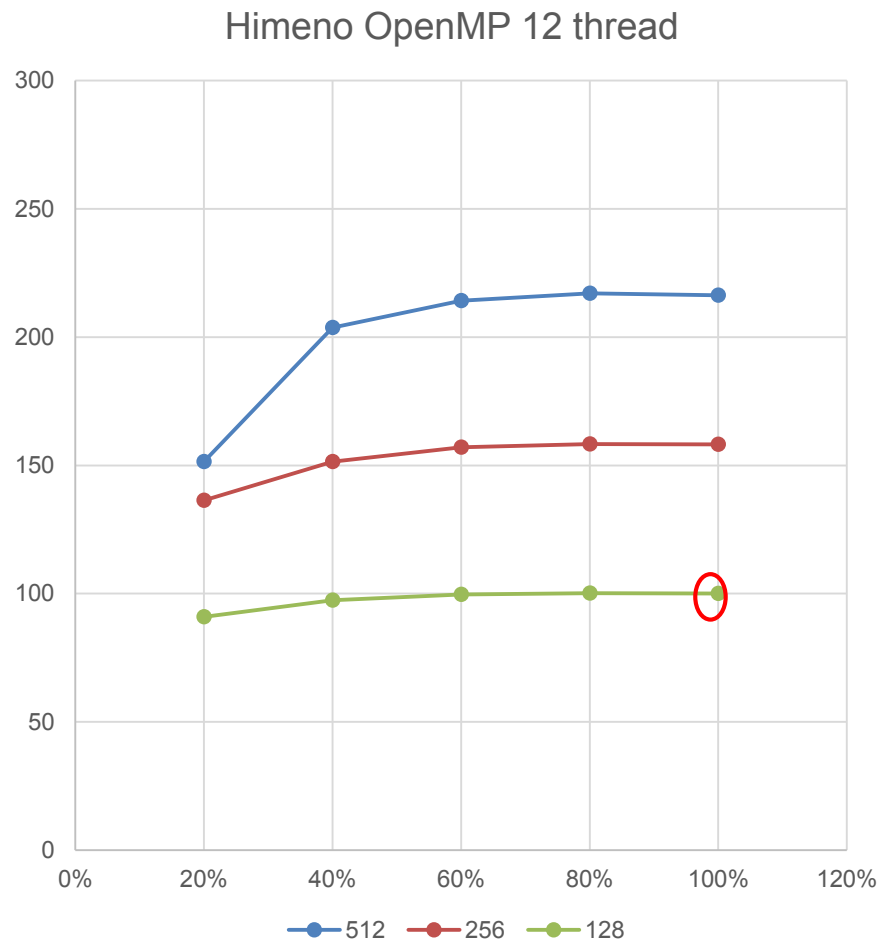


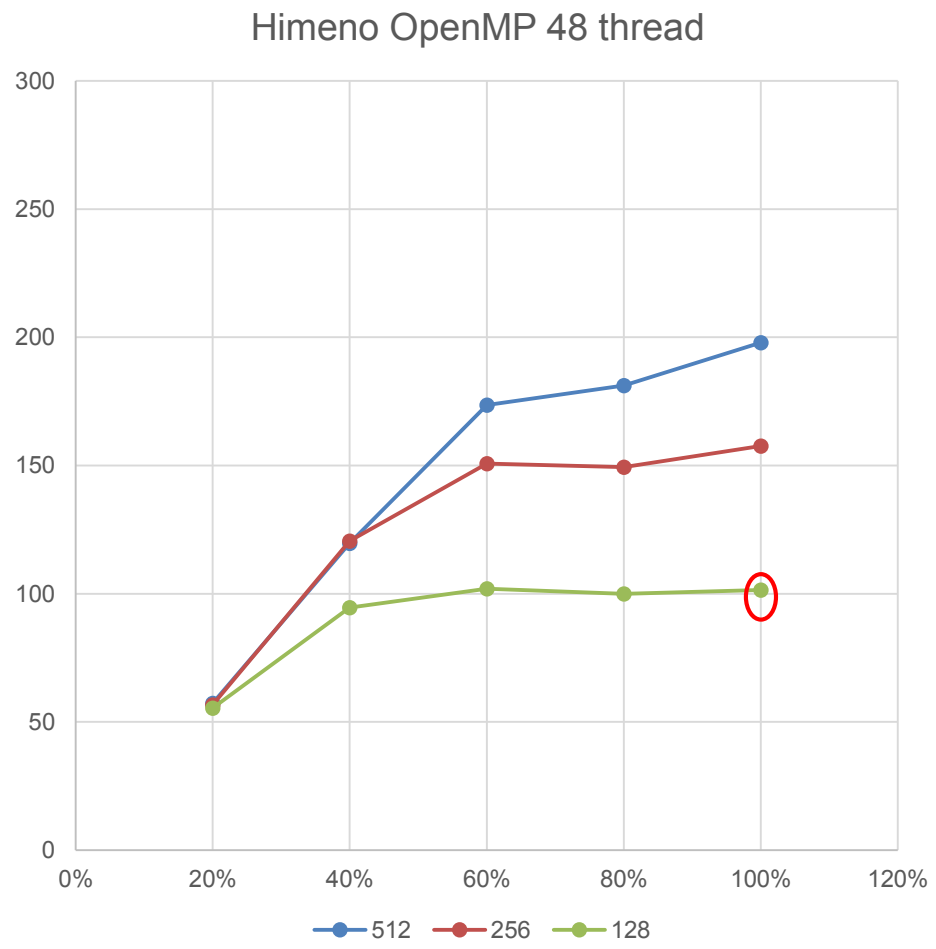



# Himeno Benchmark: 1 CMG vs. 4 CMG-SIMD Effects

FUJITSU

- Relative Performance VL128-HBM100% = 100

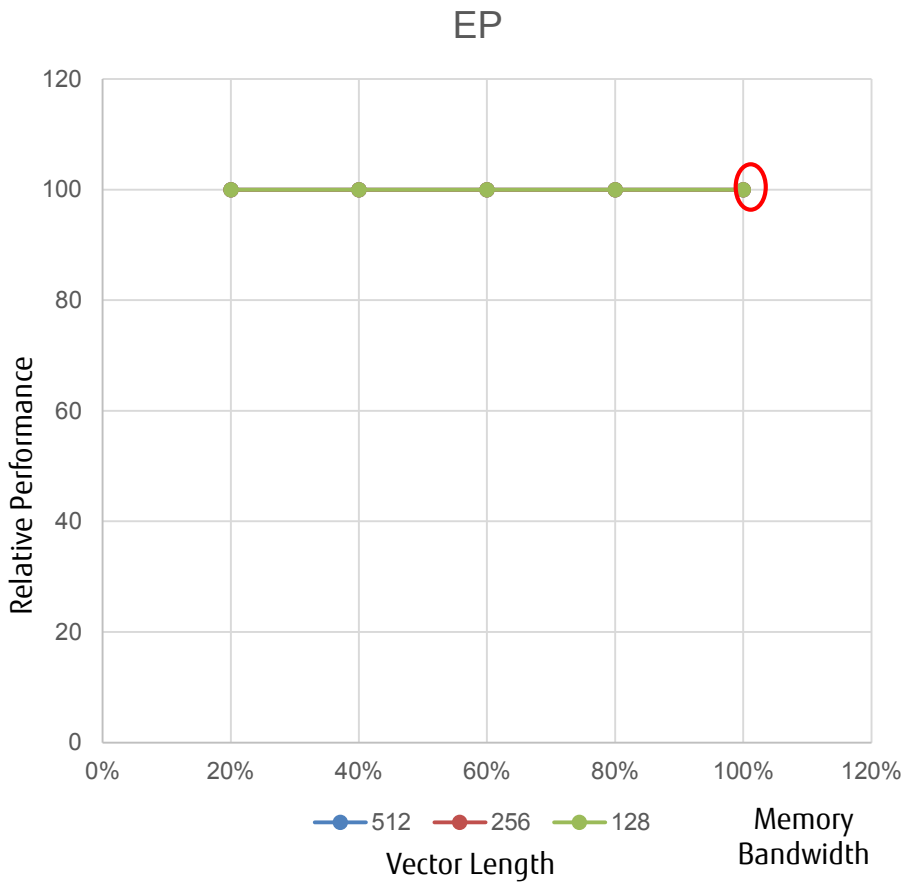


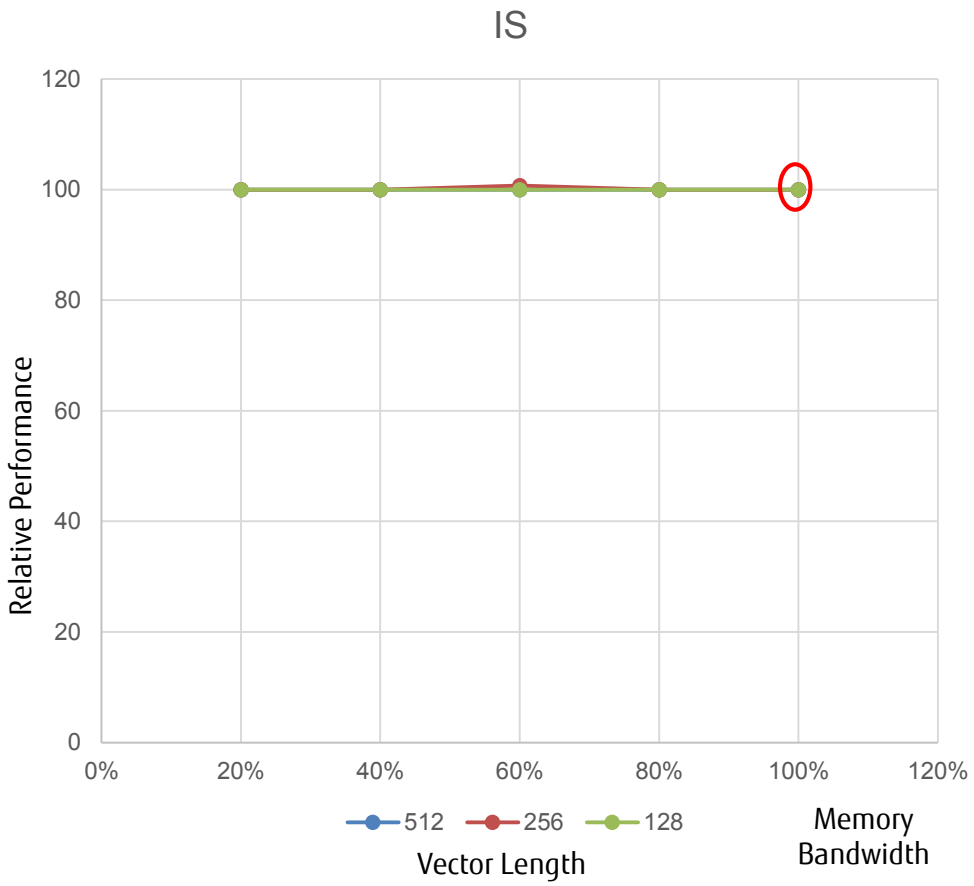



# NPB: with No SMID Effect and Compute Intensive

FUJITSU

- Relative Performance VL512-HBM100% = 100

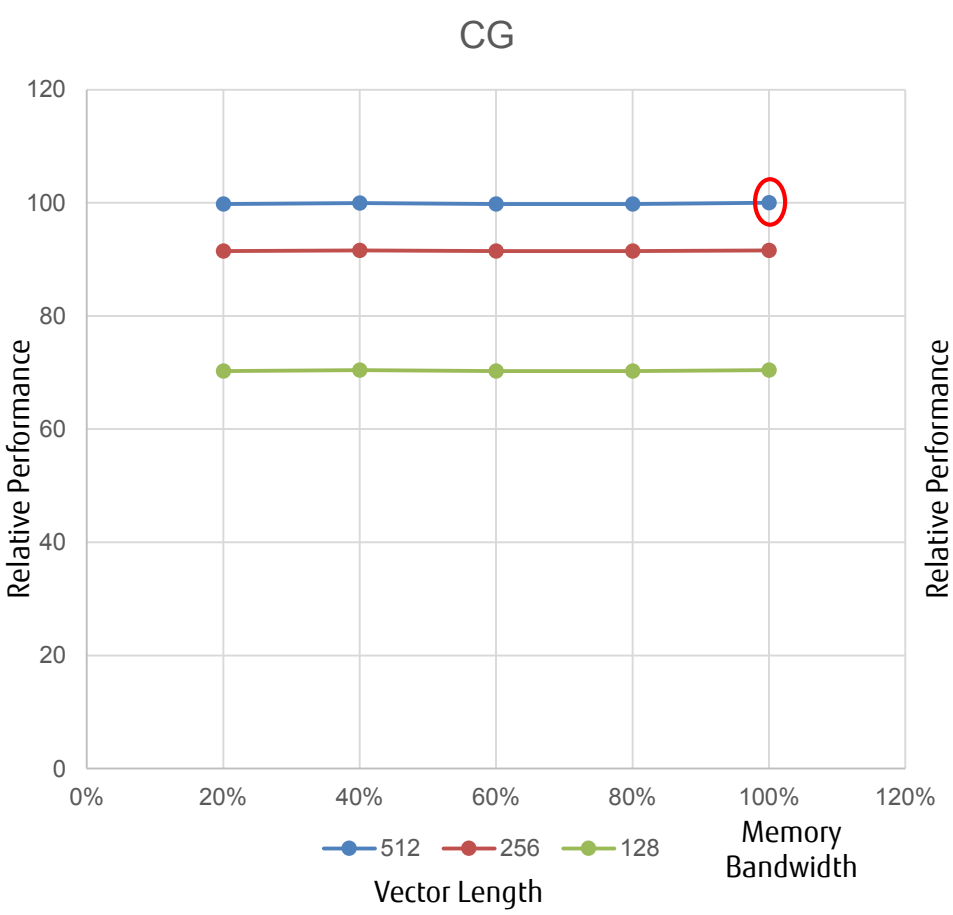


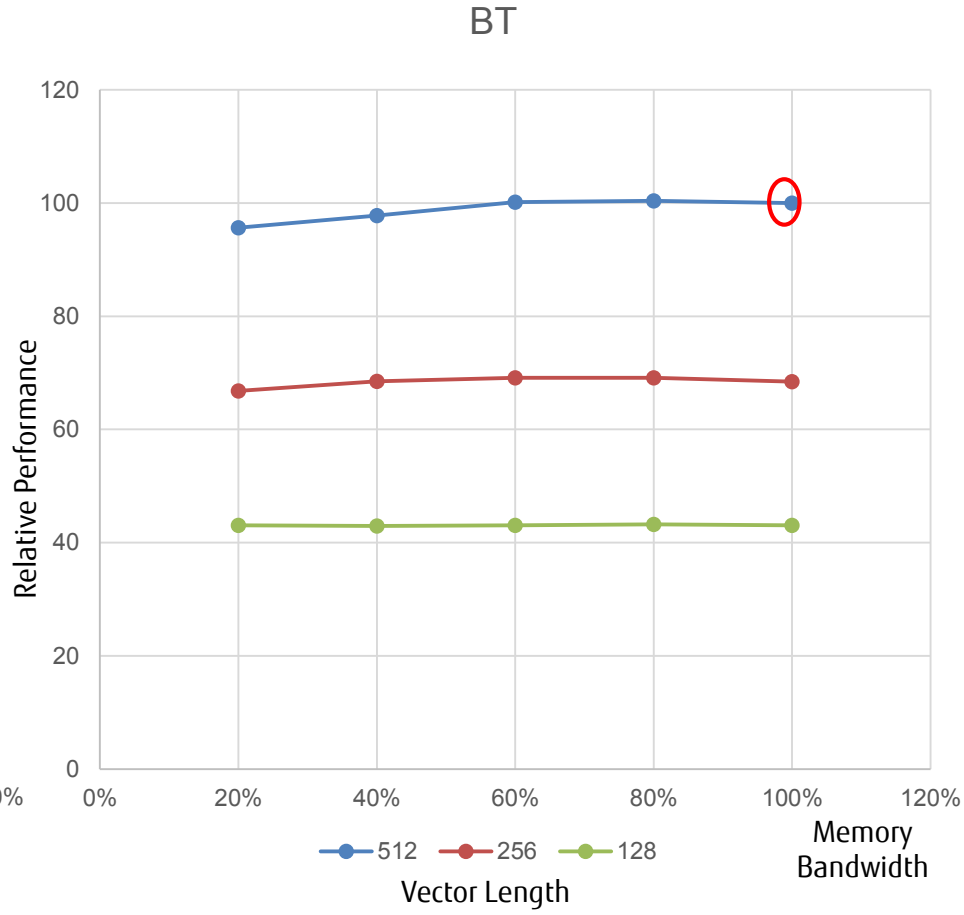



# NPB: with SMID Effect and Compute Intensive

FUJITSU

- Relative Performance VL512-HBM100% = 100

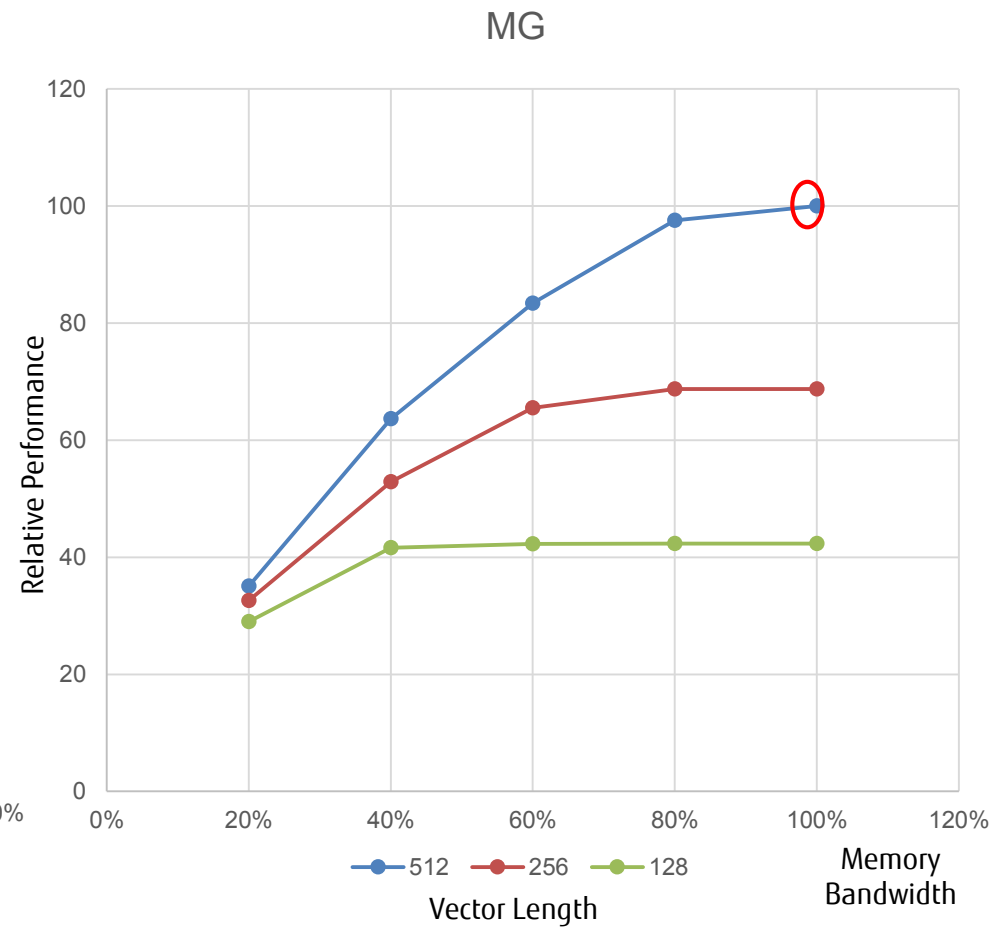


# NPB: with SMID Effect and Compute Intensive and Memory Intensive(1)

FUJITSU

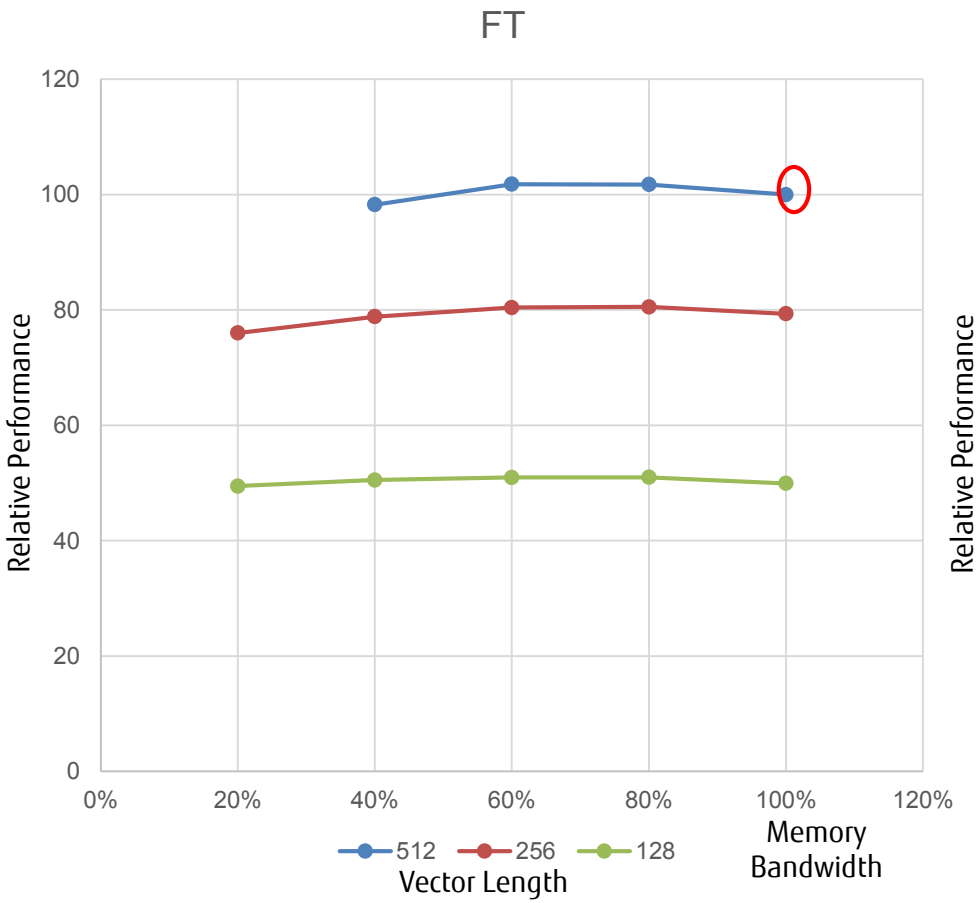
- Relative Performance VL512-HBM100% = 100

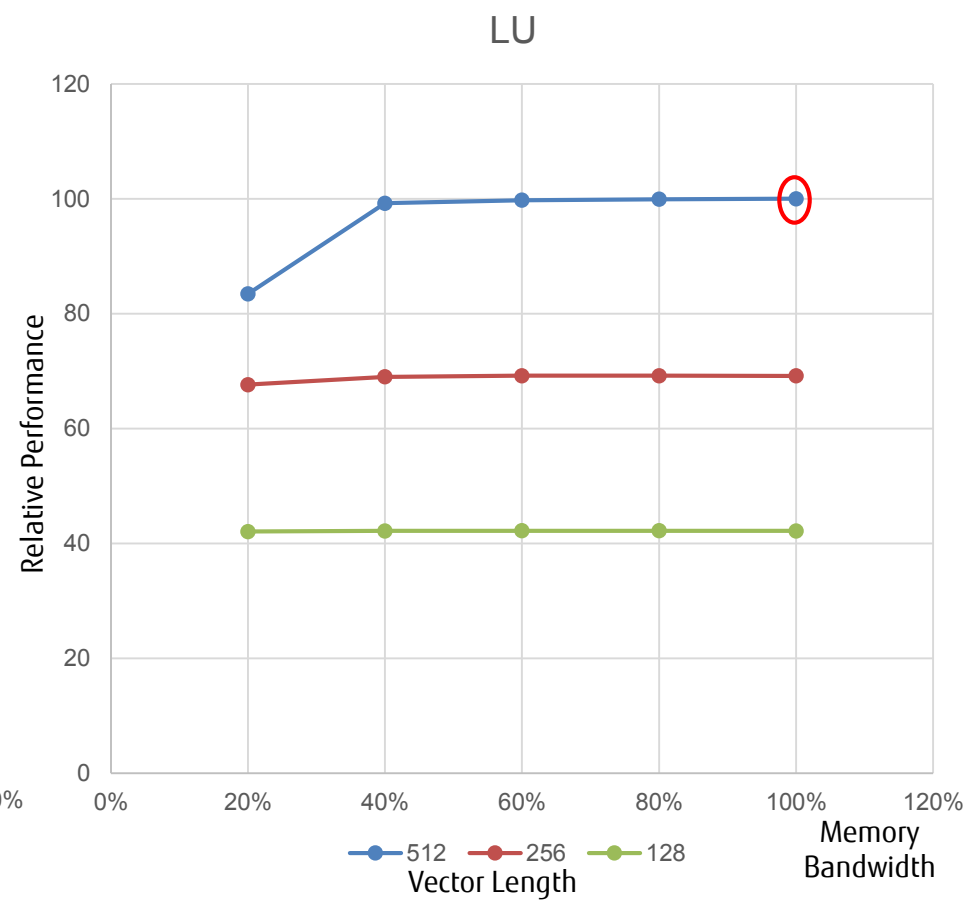



# NPB: with SMID Effect and Compute Intensive and Memory Intensive(2)

FUJITSU

- Relative Performance VL512-HBM100% = 100

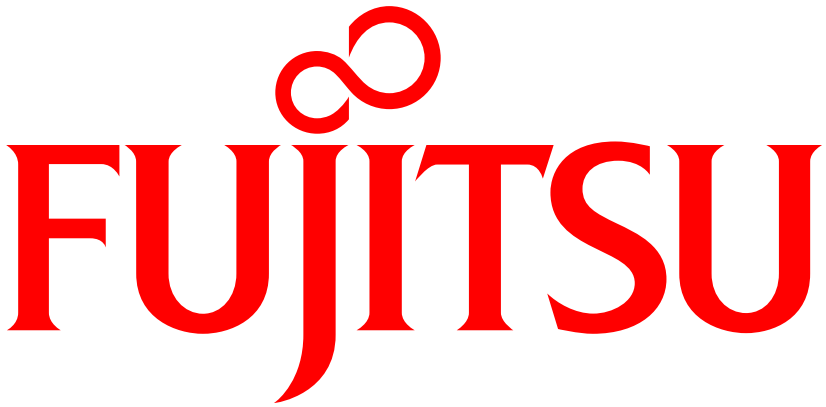


# Summary

FUJITSU

- Arm SVE provides Application Binary Portability
  - Variable Vector Length is defined at runtime
- SVE feature can be used application performance analysis
  - Reduction of vector length shows application is compute intensive or not
- Benchmark Evaluation on A64FX Enables:
  - Compute Intensive Analysis: Changing Vector Length
  - Memory Intensive Analysis:  Changing Memory Access Gap




shaping tomorrow with you