Performance evaluation and analysis of SX-Aurora TSUBASA

Kazuhiko Komatsu
Tohoku University

9 October, 2018
WSSP28

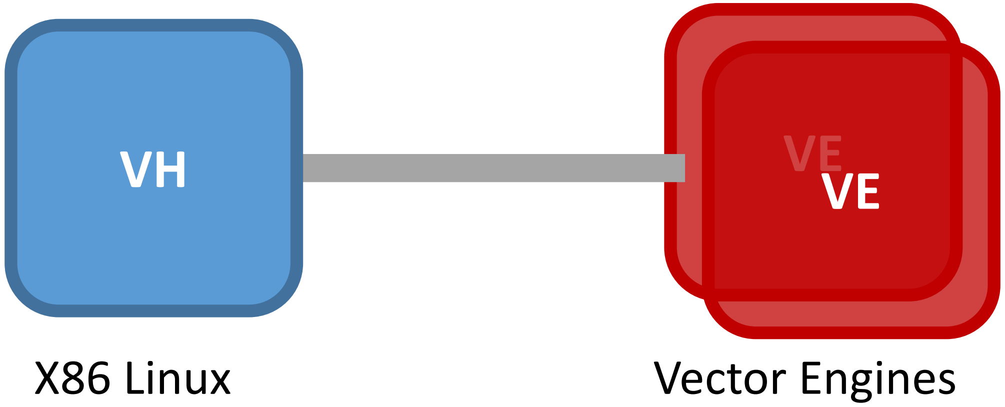
Background

- Supercomputers are important infrastructures
    - Widely used for scientific research as well as various industries
    - Top1 system reaches 122.3 Pflop/s
- Big gap between theoretical performance and sustained performance
    - **Only compute-intensive** applications stand to benefit from high peak performance
    - **Memory-intensive** applications are limited by lower memory performance

Memory performance has gained more and more attentions

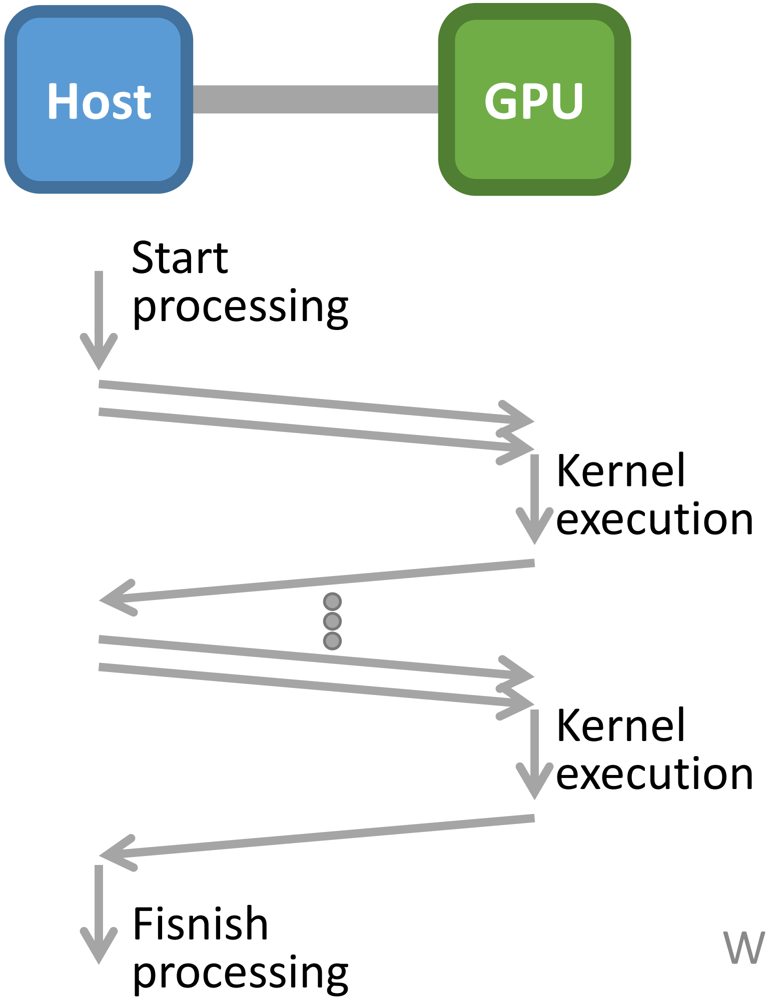
SX-Aurora TSUBASA with the world's highest memory bandwidth

- Two important concepts of its design
  - **High usability**
  - **High sustained performance**
- New architecture
  - **Vector host (VH)** is attached to **vector engines (VEs)**
    - VE is responsible for executing an entire application
    - VH is used for processing system calls invoked by the applications

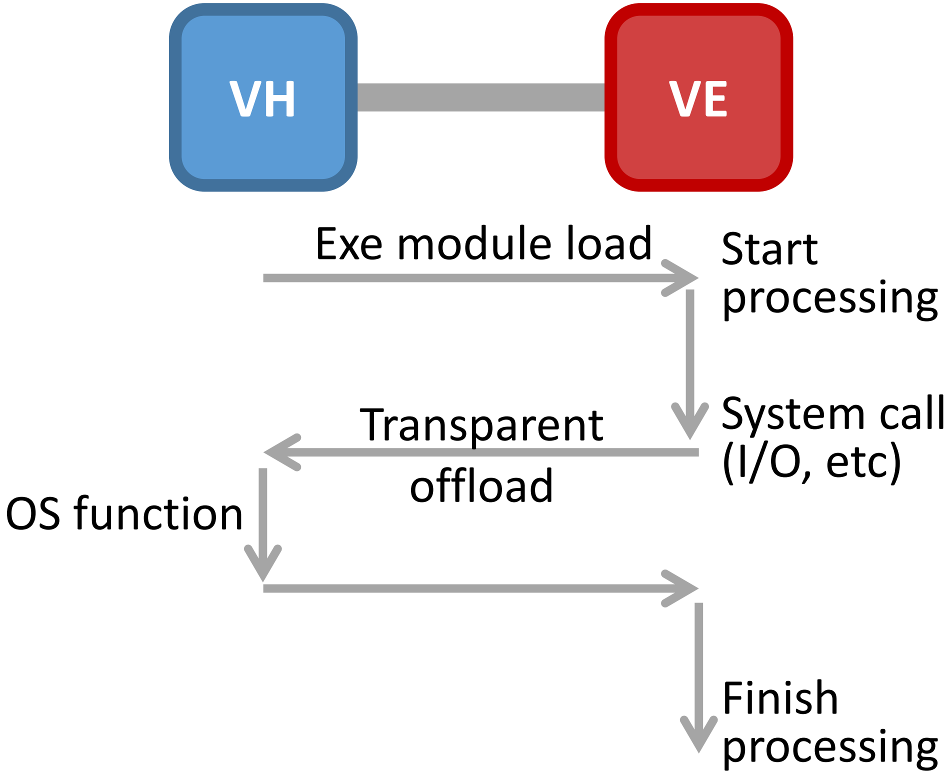


New execution model

- Conventional model



- New execution model



Highlights of the execution model

- Two advantages over conventional execution model
    - Avoid frequent data transfers between VE and VH
        - Applications are entirely executed on VE

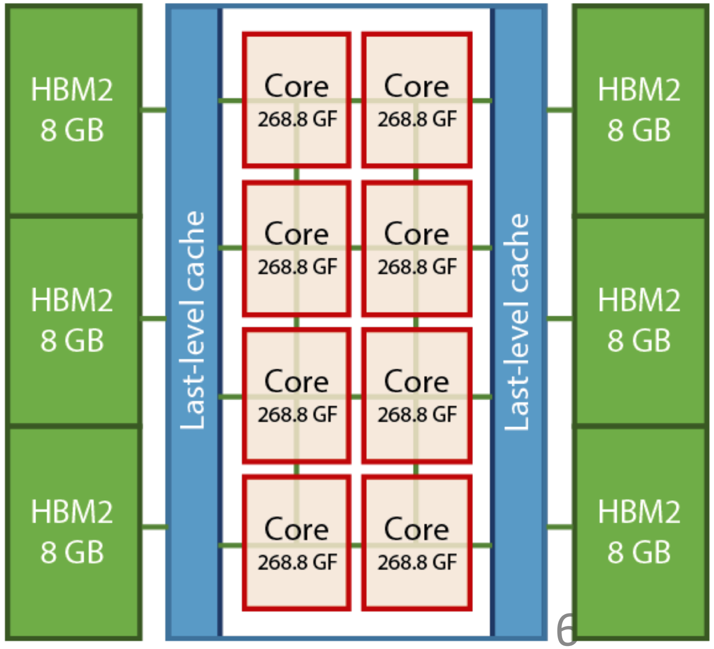
    →**High sustained performance**

    - No special programming
        - Explicit specifications of computation kernels are not necessary
        - System calls are transparently offloaded to the VH
            - Programmers do not need to care system calls

    →**High usability**

Specification of SX-Aurora TSUBASA

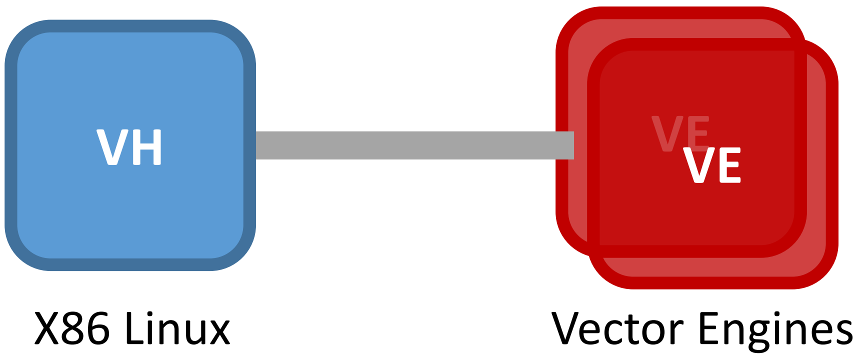
- High memory bandwidth
  - 1.22 TB/s world’s highest memory bandwidth
    - Six HBM2 memory modules integration
  - 3.0 TB/s LLC bandwidth
    - LLC is connected to cores via 2D mesh network
- High computational performance
  - 2.15 Tflop/s@1.4 GHz
    - 8 powerful vector cores
    - 16 nm FINFET process technology
    - 4.8 billion transistors
    - 14.96 mm x 33.00 mm



Block diagram of a vector processor

Performance evaluation of SX-Aurora TSUBASA

- SX-Aurora TSUBASA A300-2
  - 2x VEs Type 10B
  - 1x VH



VE	Type 10B
Frequency	1.4 GHz
Peak FP / core	268.8 GFLOPS
# cores	8
Peak DP Flops / socket	2.15 TFLOPS
Memory BW	1.2 TB/s
Memory capacity	48 GB

VH	
CPU	Intel Xeon Gold 6126
Frequency	2.60 GHz / 3.70 GHz (Turbo)
# cores	12
Mem BW	128 GB/s
Mem Capacity	96 GB
Mem config	DDR4-2666 DIMM 16GB x 6

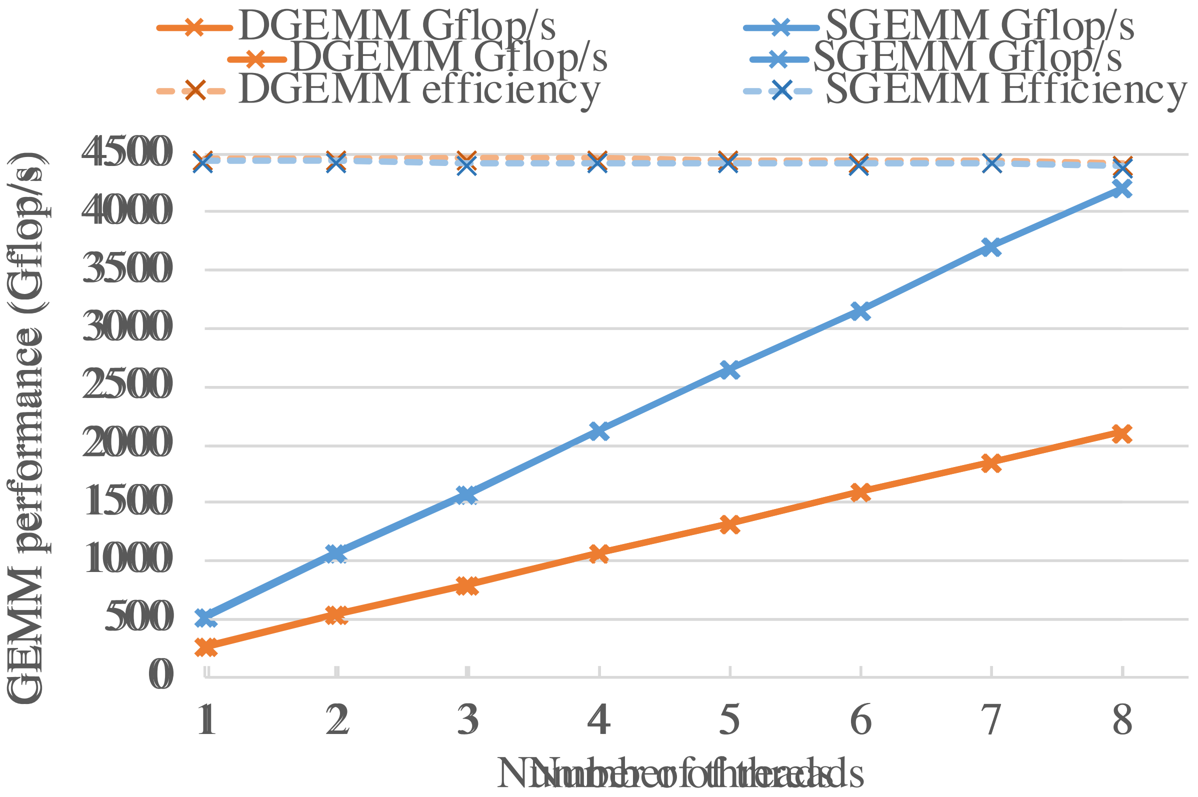
Experimental environments

Processor	SX-Aurora Type 10B	Xeon Gold 6126	SX-ACE	Tesla V100	Xeon Phi KNL 7290
Frequency	1.4 GHz	2.6 GHz	1.0 GHz	1.245 GHz	1.5 GHz
# of cores	8	12	4	5120	72
DP flop/s (SP flop/s)	2.15 T (4.30 T)	998.4 GF (1996.8 GF)	256 GF	**7 TF** **(14 TF)**	3.456 TF (6.912 TF)
Memory subsystem	**HBM2 x6**	DDR4 x6ch	DDR3 x16ch	HBM2	MCDRAM DDR4
Memory BW	**1.22 TB/s**	128 GB/s	256 GB/s	900 GB/s	450+ GB/s 115.2 GB/s
Memory capacity	48 GB	96 GB	64 GB	16 GB	16 GB 96 GB
LLC BW	2.66 TB/s	N/A	1.0 TB/s	N/A	N/A
LLC capacity	16 MB shared	19.25 MB shared	1 MB private	6 MB shared	1 MB shared by 2 cores

Applications used for evaluation

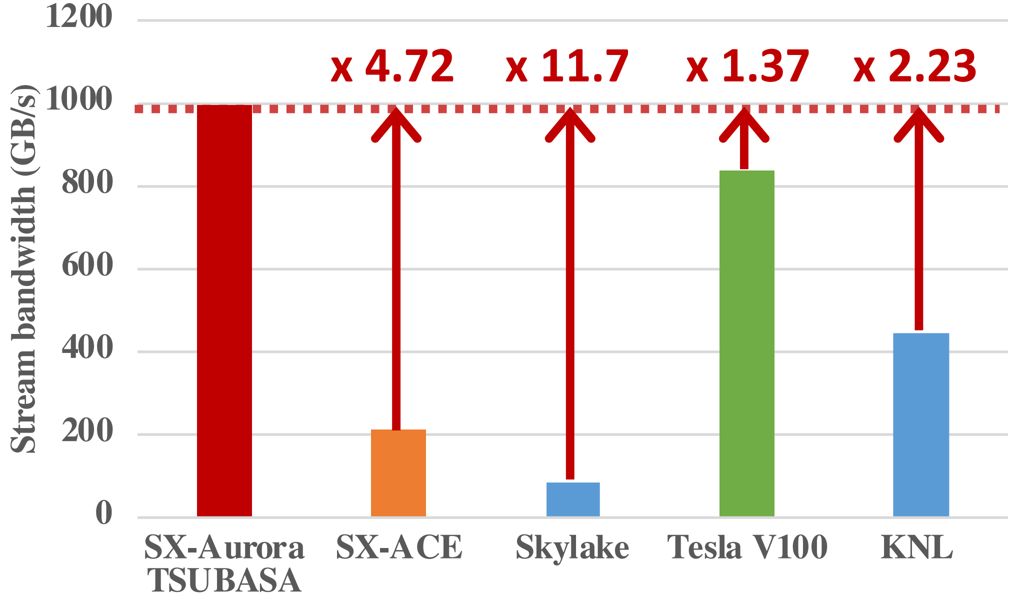
- SGEMM/DGEMM
  - Matrix-matrix multiplications to evaluate the Peak flop/s
- Stream benchmark
  - Simple kernels (copy, scale, add, triad) to measure sustained memory performance
- Himeno benchmark
  - Jacobi kernels with a 19-point stencil as a memory-intensive kernels
- Tohoku univ's kernels
  - Kernels of practical applications of Tohoku univ in Earthquake, CFD, Electromagnetic
- Microbenchmark for offload evaluation
  - Mixture with vector-friendly jacobi kernels and I/O kernels

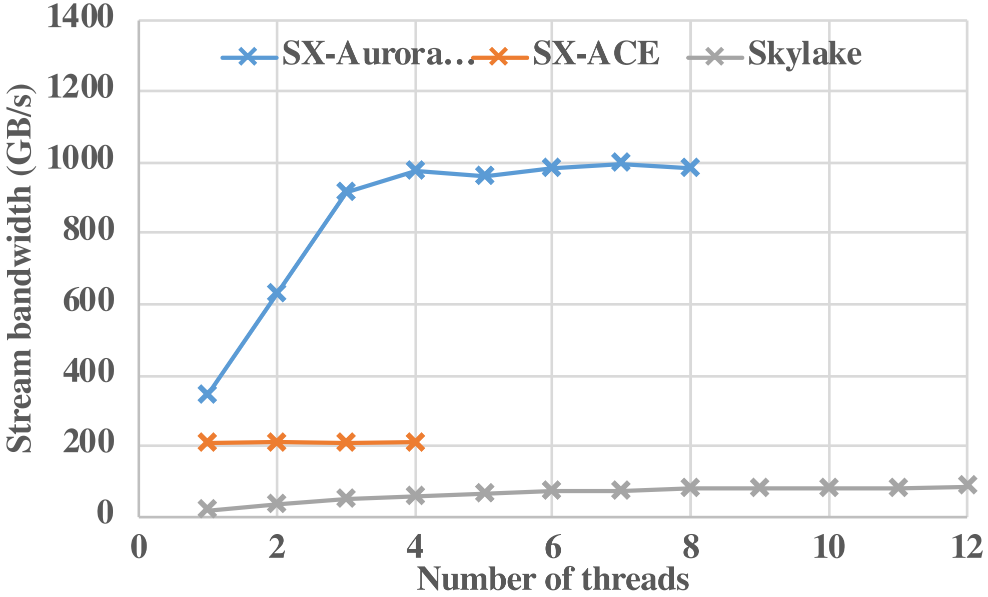
SGEMM/DGEMM Performance



- High scalability up to 8 threads
  - High vectorization ratio 99.36%, good vector length 253.8
- High efficiency and achieve almost ideal performance
  - Efficiecy 97.8~99.2%

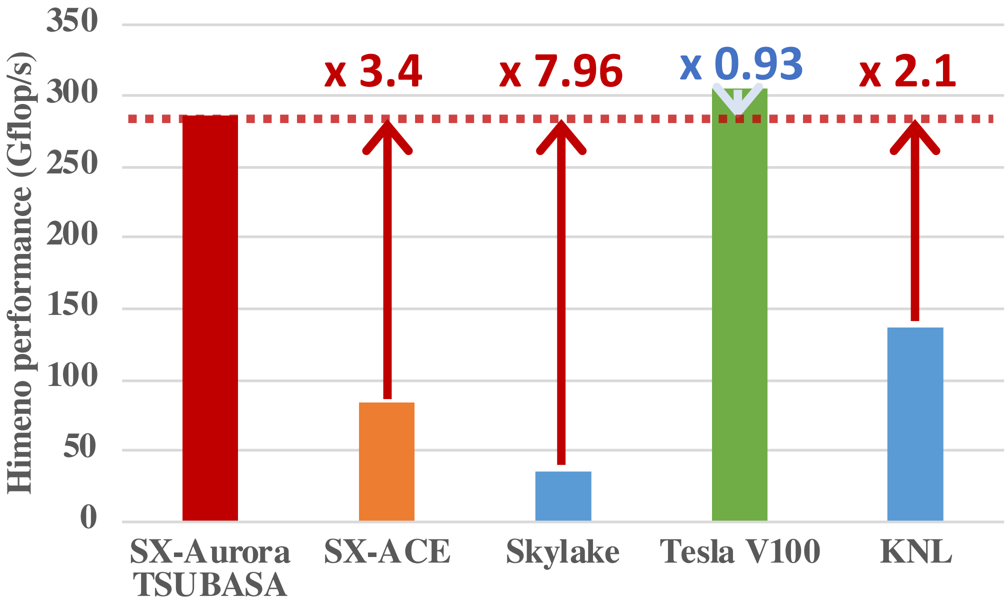
Memory performance(Stream Triad)

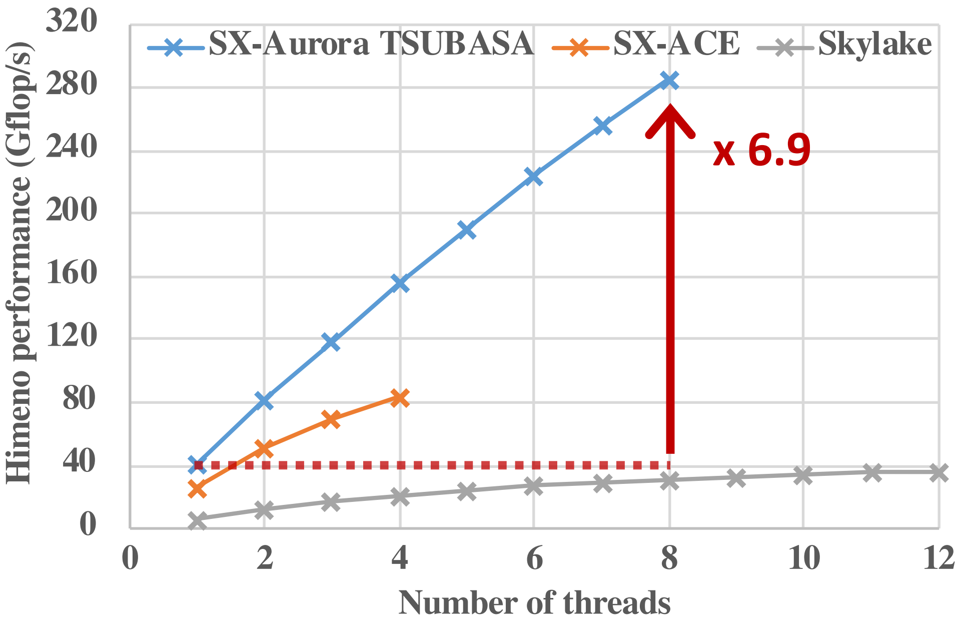




- High sustained memory bandwidth of SX-Aurora TSUBASA
  - Efficiency: Aurora 79%, ACE 83%, Skylake 66%, V100 81%
- Scalability
  - Saturated even when the number of cores is 3 or 4

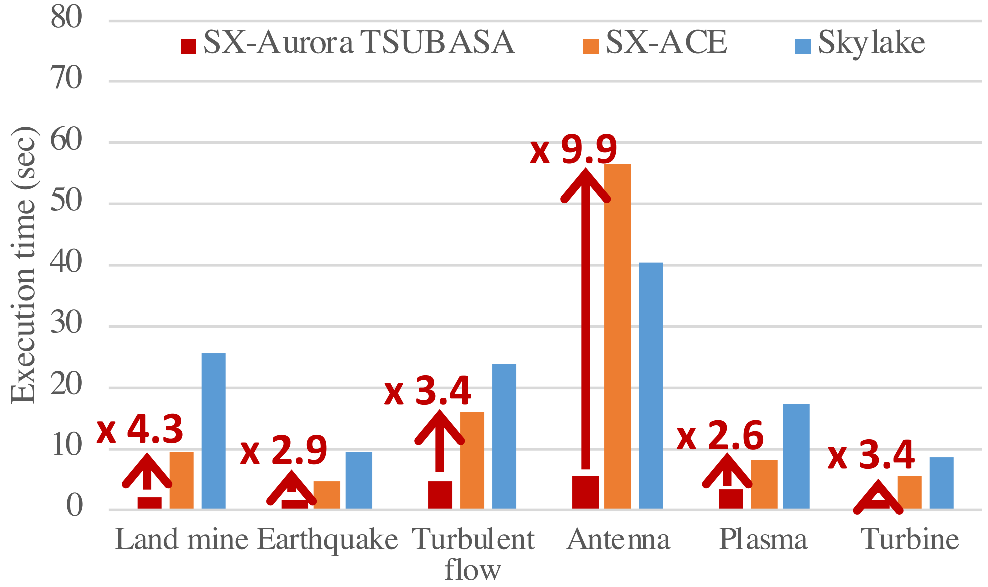
Himeno (Jacobi) performance

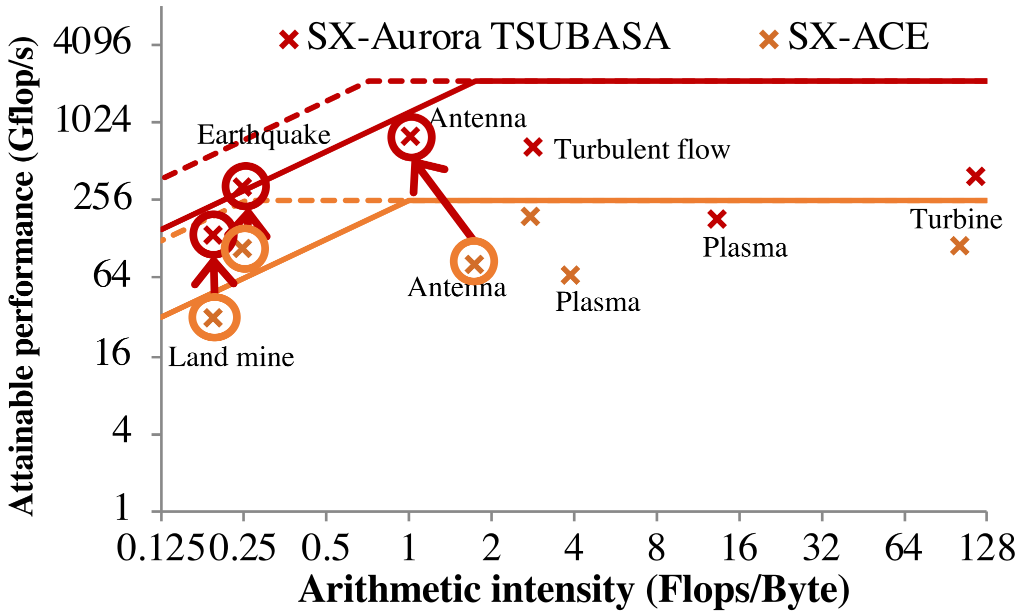




- Higher performance… except GPU
    - Vector reduction becomes bottleneck due to copy among vector pipes
- Nice thread scalability
    - 6.9x speedup in 8 threads => 86% parallel efficiency (OpenMP overhead?)

Application kernel performance





- SX-Aurora TSUBASA could achieve high performance
  - Plasma, Turbine => Indirect access, memory latency-bound
  - Antenna => computation-bound to memory BW-bound
  - Land mine, Earthqauke, Turbulent flow =>memory or LLC BW-bound

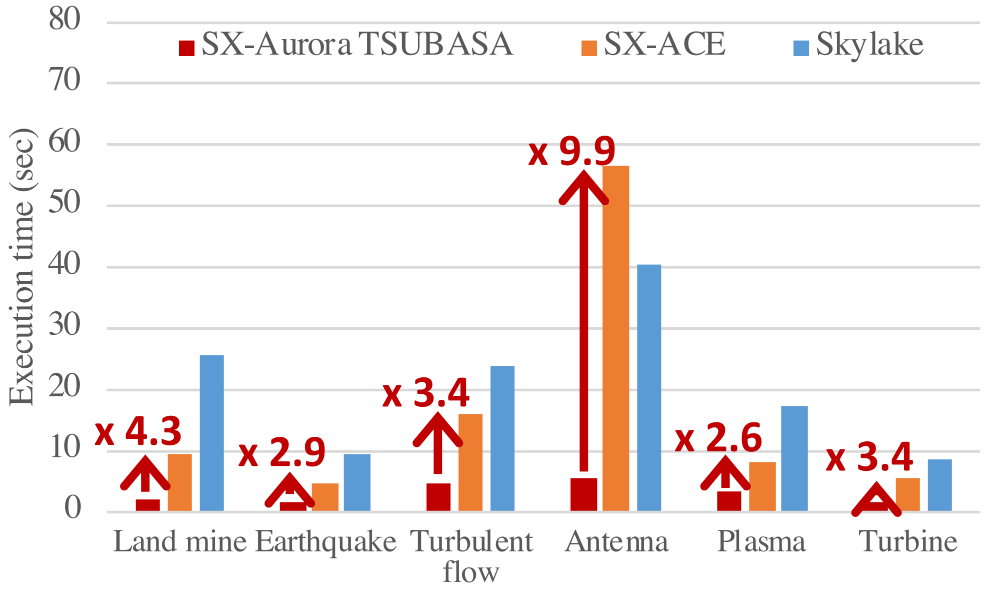
Memory bound? or LLC bound?

- Further analysis using 4 types of Bytes/Flop ratio
  - **Memory B/F** = (memory BW) / (peak performance)
  - **LLC B/F** = (LLC BW) / (peak performance)
  - **Code B/F** = (necessary data in Byte) / (# FP operations)
  - **Actual B/F** = (# block memory access) * (block size) / (# FP operations)

B/F ratio	Actual < Memory	Memory > Actual
Code < LLC	Computation-bound	Memory BW-bound
Code > LLC	LLC BW-bound	Memory or LLC bound *

  - Code B/F > Actual B/F * LLC BW / Memory BW => **LLC bound**
  - Code B/F < Actual B/F * LLC BW / Memory BW => **memory bound**

Application kernel performance

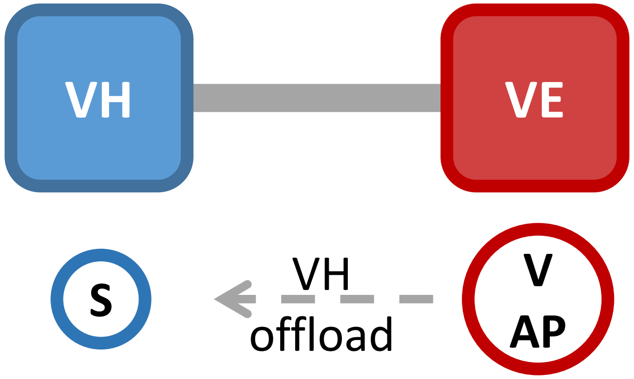


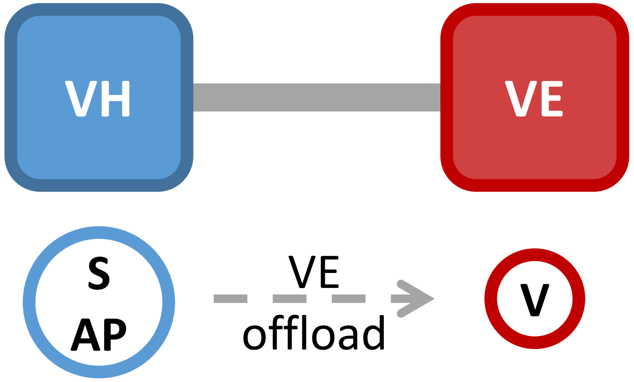
B/F ratio	Actual < Memory	Memory > Actual
Code < LLC	Computation-bound	Memory BW-bound
Code > LLC	LLC BW-bound	Memory or LLC bound *

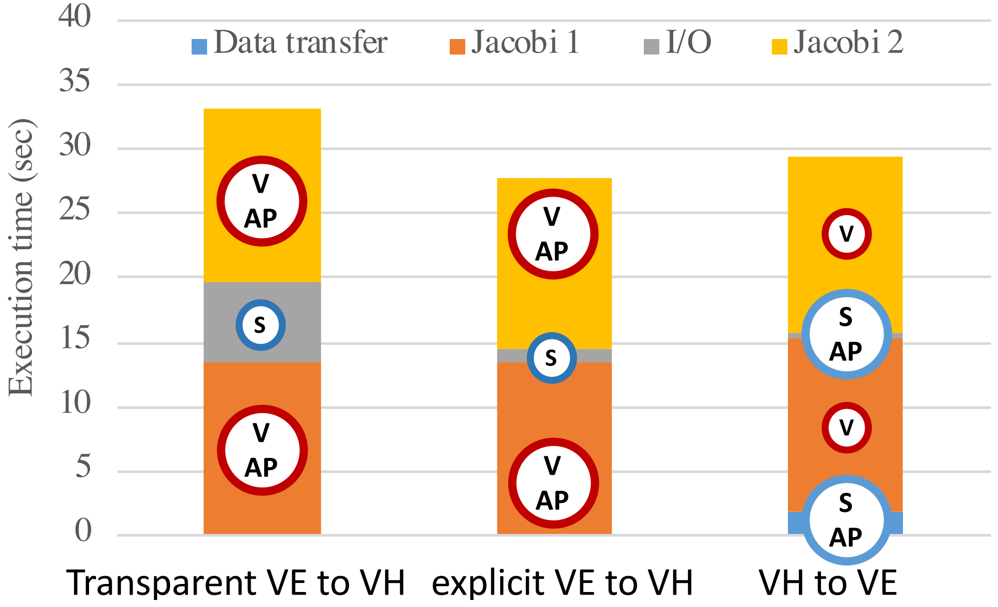
- Land mine => LLC-bound
- Earthqauke => LLC-bound
- Turbulent flow => memory BW-bound
- Antenna => memory BW-bound

Evaluation of the execution model

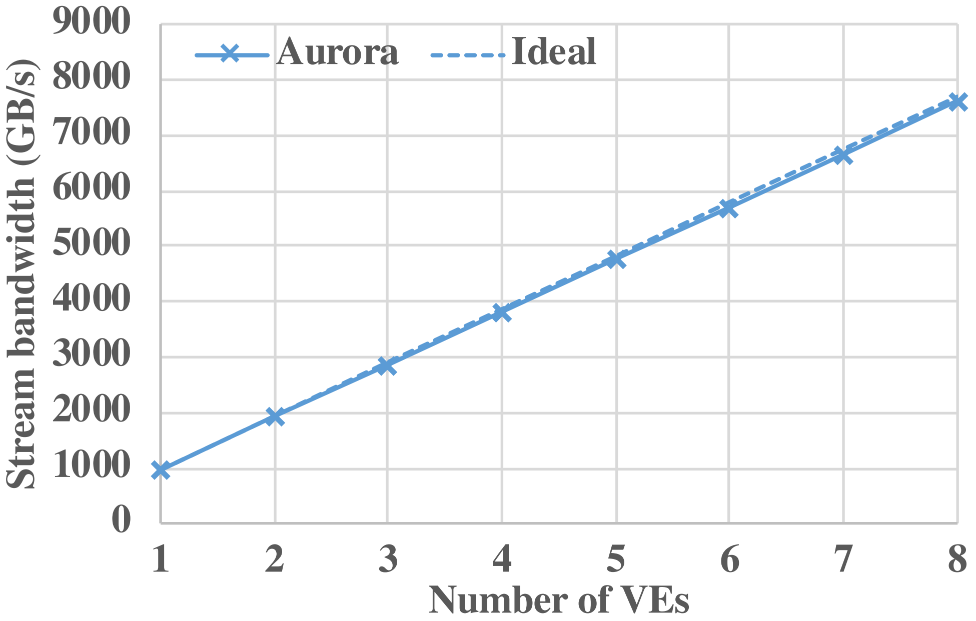
- (Transparent/Explicit) Offload from VE to VH
- Offload from VH to VE

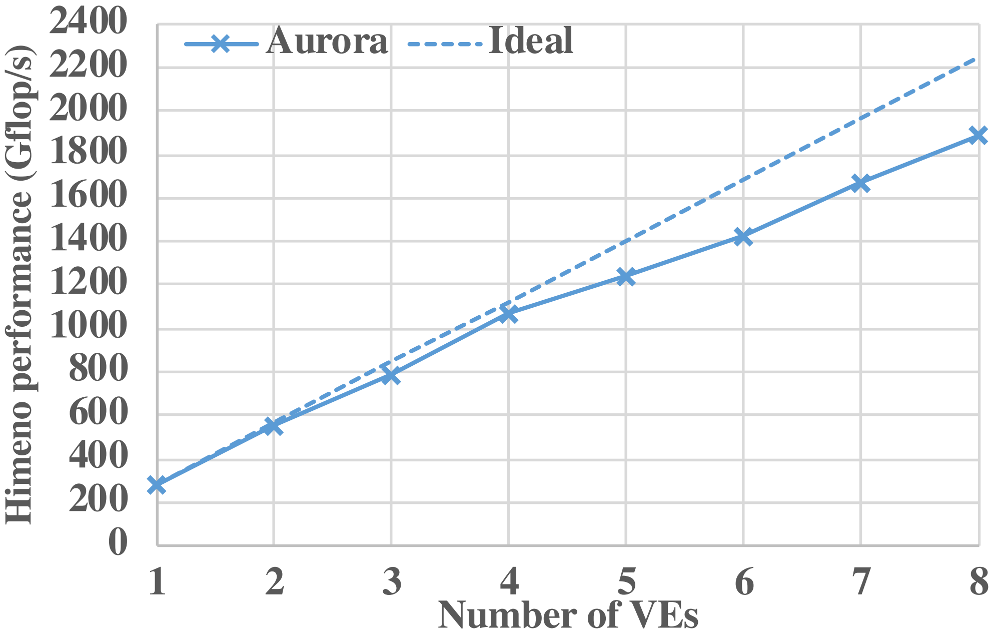






Multi-VE performance on A300-8





- Stream VE-level scalability
  - Almost ideal scalability up to 8 VEs
- Himeno VE-level scalability
  - Good scalability up to 4VEs
  - Lack of vector lengths when more than 5VEs
    - Problem size is too small

Conclusions

- Performance evaluation and analysis of SX-Aurora TSUBASA
  - Standard benchmark programs

  → **High potential of compute and memory performances**
  - Kernels of practical applications

  → **High memory performance leads high sustained performance**
  - Microbenchmark

  → **Effectiveness of a new execution model**