NATIONAL RESEARCH UNIVERSITY

Суперкомпьютерные дни в России 2025

National Research University Higher School of Economics

Supercomputer Atomistic Modelling and Multi-scale Analysis

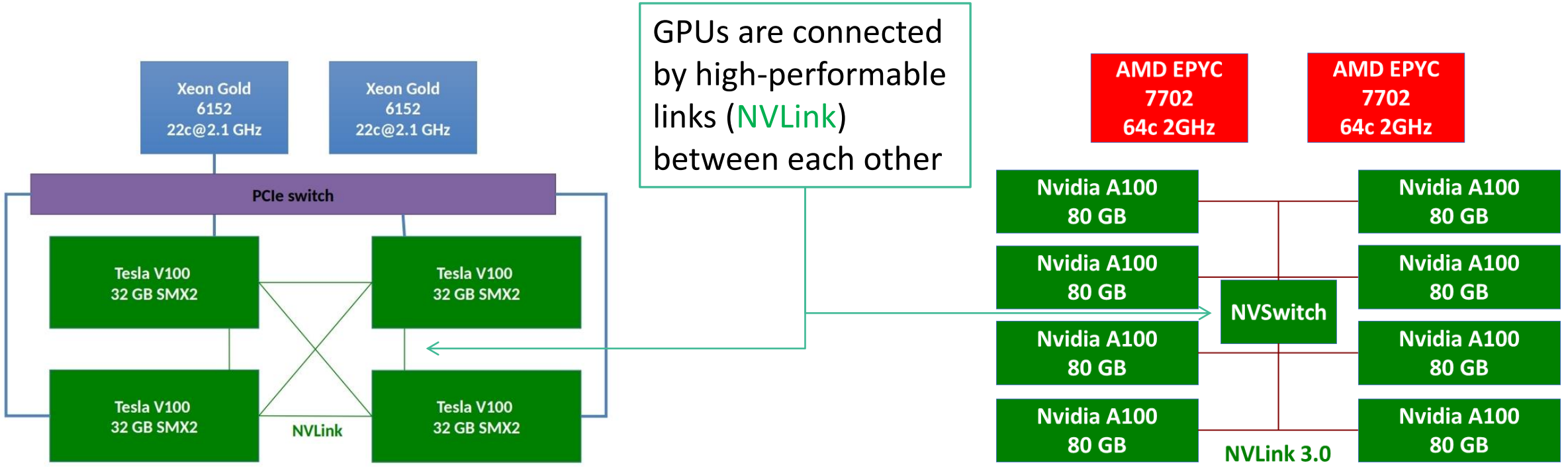
GEMM ALGORITHM FOR MULTI-GPU PLATFORMS WITH REGULAR UNEVEN DATA TRANSFER LINKS

Choi Yea Rem, Malkovsky S., Stegailov V.

Moscow, 2025

MULTI-GPU GEMM ALGORITHM

Testing platforms

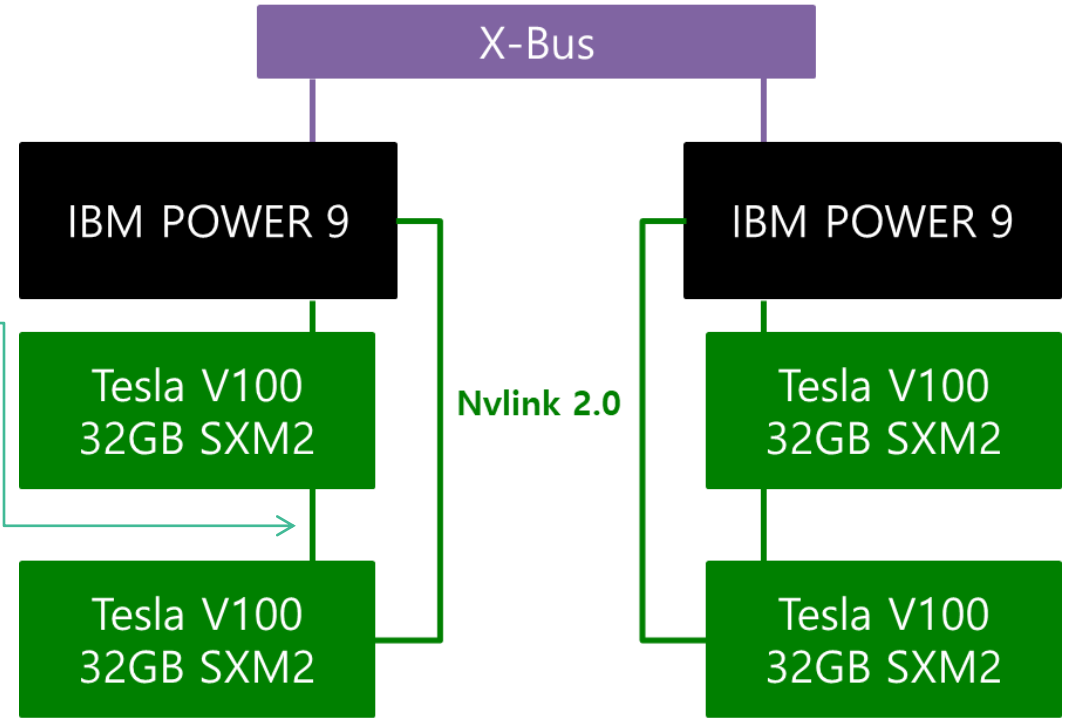


The topology of a node with 4 Nvidia V100 GPUs in the «cHARISMa» supercomputer

The topology of a node with 8 Nvidia A100 GPUs in the «cHARISMa» supercomputer

Testing platforms

2 V100 GPUs and one IBM POWER 9 CPU are connected by high-performable links (NVLink)



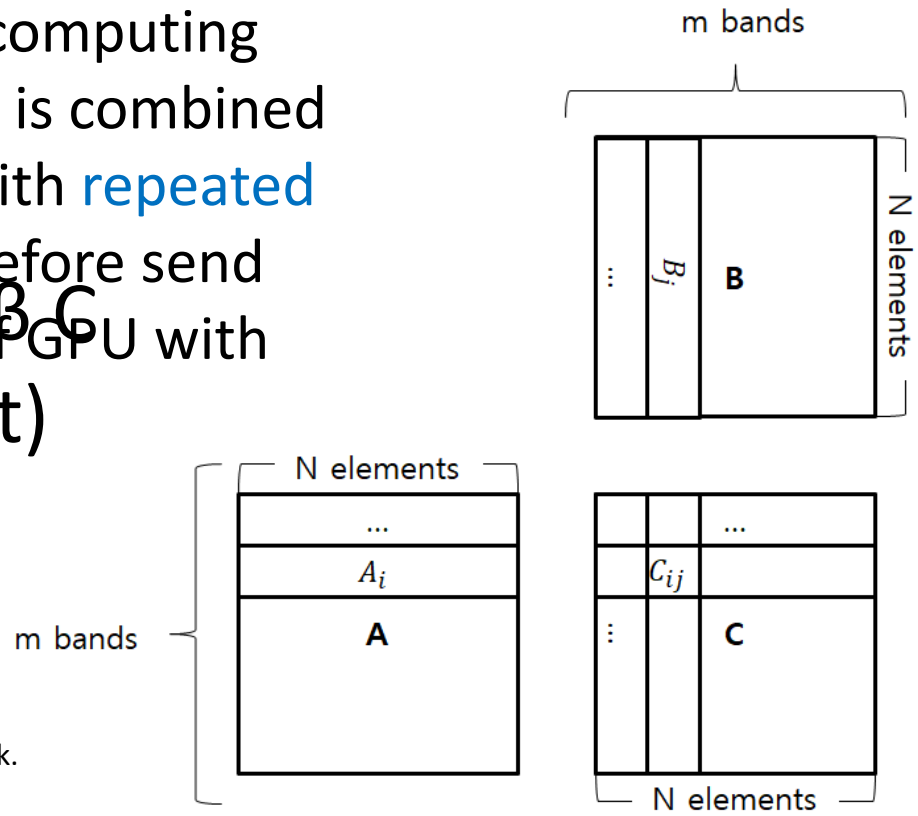
The topology of IBM Power System AC922 8335-GTH with 4 V100 GPU

Strategy of the developed GEMM algorithm

To reduce the amount and frequency of data transfer in the computing process bands of matrix C is combined by tiles on worker GPU with repeated usage of band A data before send result to the memory of GPU with

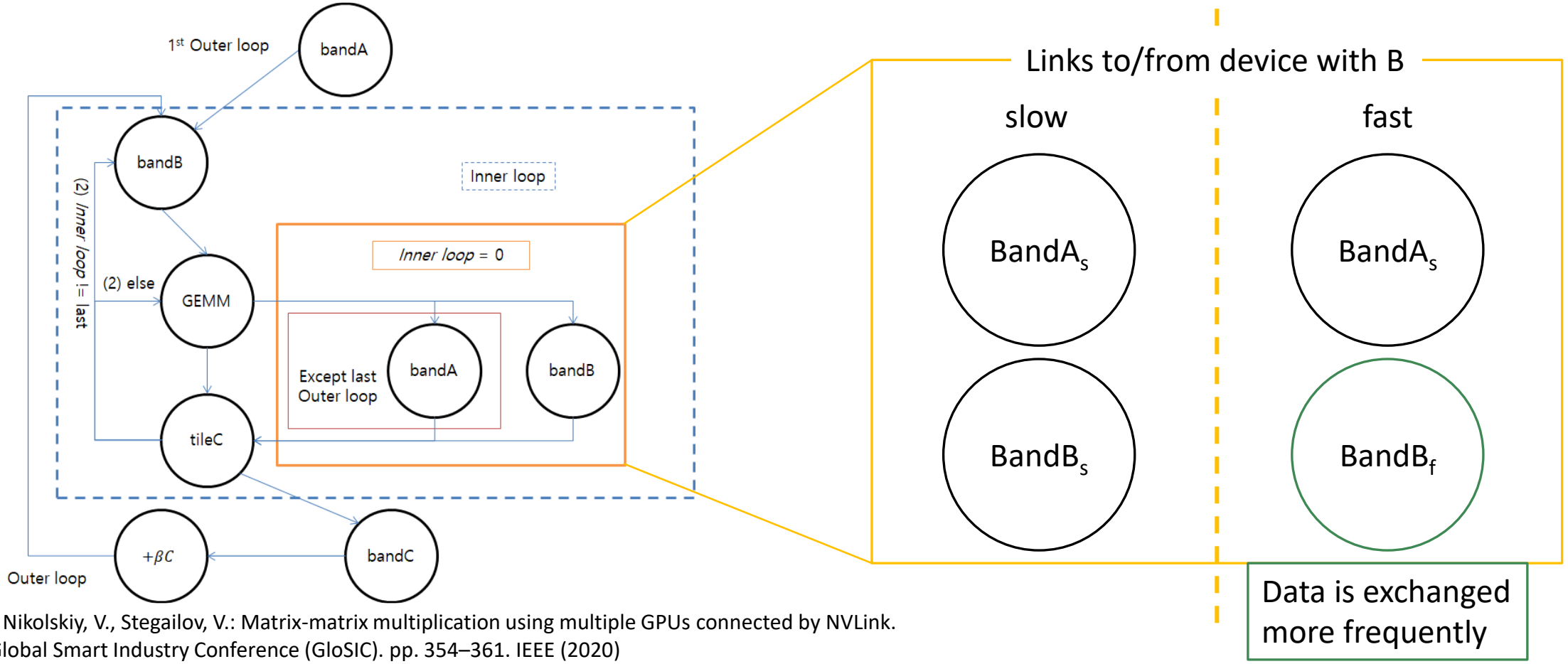
$$C = \alpha (AB) + \beta C$$

gemm (α, β – const)



Choi, Y.R., Nikolskiy, V., Stegailov, V.: Matrix-matrix multiplication using multiple GPUs connected by NVLink. In: 2020 Global Smart Industry Conference (GloSIC). pp. 354–361. IEEE (2020)

Scheme of the developed GEMM algorithm on multiple GPUs



Choi, Y.R., Nikolskiy, V., Stegailov, V.: Matrix-matrix multiplication using multiple GPUs connected by NVLink. In: 2020 Global Smart Industry Conference (GloSIC). pp. 354–361. IEEE (2020)

COMPUTING EXPERIMENTS: SETTINGS

Testing platform parameters

Hardware Parameters	IBM POWER 4 x V100	cHARISMa 4 x V100	cHARISMa 8 x A100
Peak FP32 performance (TFLOPS)	14.899	14.899	-
Real FP32 performance (TFLOPS)	14.8	14.8	-
Peak TF32 performance (TFLOPS)	-	-	156
Real TF32 performance (TFLOPS)	-	-	124
Peak GPU memory bandwidth (GB/s)	900	900	2039
Peak GPU-GPU bandwidth (GB/s)	75 64	50 15.754	300 31.508
Real GPU-GPU bandwidth (GB/s)	72.68 33.24	48.33 9.74	281 17.42

Theoretical tile sizes for platforms with two different transfer bandwidths

$$\begin{cases} N_{i_s} > 2BW_{math}/BW_{transfer_s} \\ N_{i_f} > 2BW_{math}/BW_{transfer_f} \end{cases}$$

Expected in the idle condition if data transfer and computation overlapping is not affected by the tile sizes

Unequal tile sizes causes computing load difference within one kernel execution

$$\begin{cases} FLOPS_f = NN_{i_s}N_{i_f}, \\ FLOPS_s = NN_{i_s}^2, \end{cases} \quad BW_{transfer_f}$$

$$T_{math} = FLOPS/BW_{math}$$

$$T_{transfer} = 4N_iN/BW_{transfer}$$

$$\begin{cases} N_{i_s} > 2BW_{math}/BW_{transfer_s}, \\ N_{i_s} > 2BW_{math}/BW_{transfer_f}. \end{cases}$$

However, choosing sufficiently small N_{i_f} allows more stable asynchronous execution

Choi Y. R., Nikolskiy V., Stegailov V. Tuning of a Matrix-Matrix Multiplication Algorithm for Several GPUs Connected by Fast Communication Links, in: *Parallel Computational Technologies: 16th International Conference, PCT 2022, Dubna, Russia, March 29–31, 2022, Revised Selected Papers*. Springer, 2022. Ch. 12. P. 158-171.

Theoretical tile sizes for IBM Power System

The idle case

$$N_{i_s} > 2BW_{math}/BW_{transfer_s}$$
$$N_{i_f} > 2BW_{math}/BW_{transfer_f}$$

$$N_{i_s} > 891$$
$$N_{i_f} > 446$$

For the algorithm with one tile size

$$N_i > 2BW_{math}\left(\frac{NumGPUs_1}{BW_{transfer_1}} + \frac{NumGPUs_2}{BW_{transfer_2}}\right)$$

$$N_i > 2189$$

COMPUTING EXPERIMENTS: RESULTS

Theoretical tile sizes and performance of the algorithm on testing platforms

Testing Platforms	IBM POWER 4 x V100	cHARISMa 4 x V100	cHARISMa 8 x A100
Tile sizes, idle case	512	1024	1024
	1024	4096	8192
Performance (TFLOPS)	16.97	9.97	15.56
(% from peak)	28.48	16.76	1.57
Tile size, predicted	4096	8192	86301 (4096)
Performance (TFLOPS)	38.39	32.76	- (53.97)
(% from peak)	64.42	54.97	- (5.44)

Best empirical tile sizes and performance of the algorithm on testing platforms

IBM POWER 4 x V100

$N = 32768$
$N_{i_s} = 4096$
$N_{i_f} = 2048$
TFLOPS = 45.48
% = 76.82

cHARISMa 4 x V100

$N = 32768$
$N_{i_s} = 8192$
$N_{i_f} = 8192$
TFLOPS = 32.76
% = 54.97

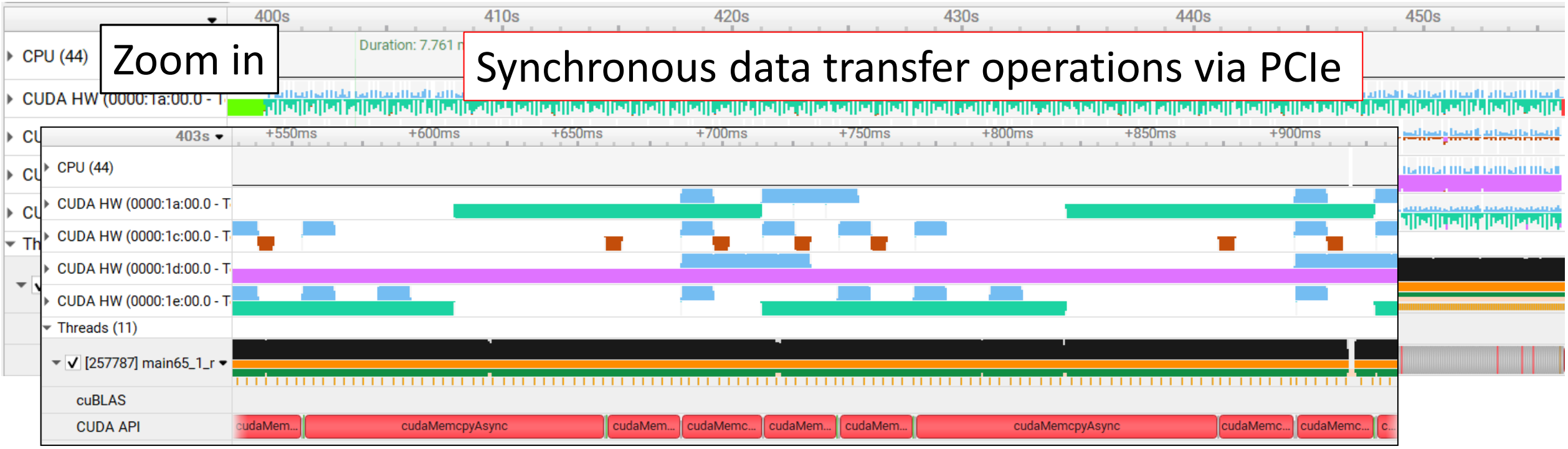
cHARISMa 8 x A100

Performance remains suboptimal compared to single A100 execution despite tile size optimization attempts

It can show over 450 TFLOPS on 8 x A100 with all-to-all NVLink topology

Choi Y. R., Stegailov V. Multi-GPU GEMM Algorithm Performance Analysis for Nvidia and AMD GPUs Connected by NVLink and PCIe. In: *22nd International Conference, MMST 2022, Nizhny Novgorod, Russia, November 14–17, 2022, Revised Selected Papers*. Springer, 2022. Ch. 23. P. 281-292.

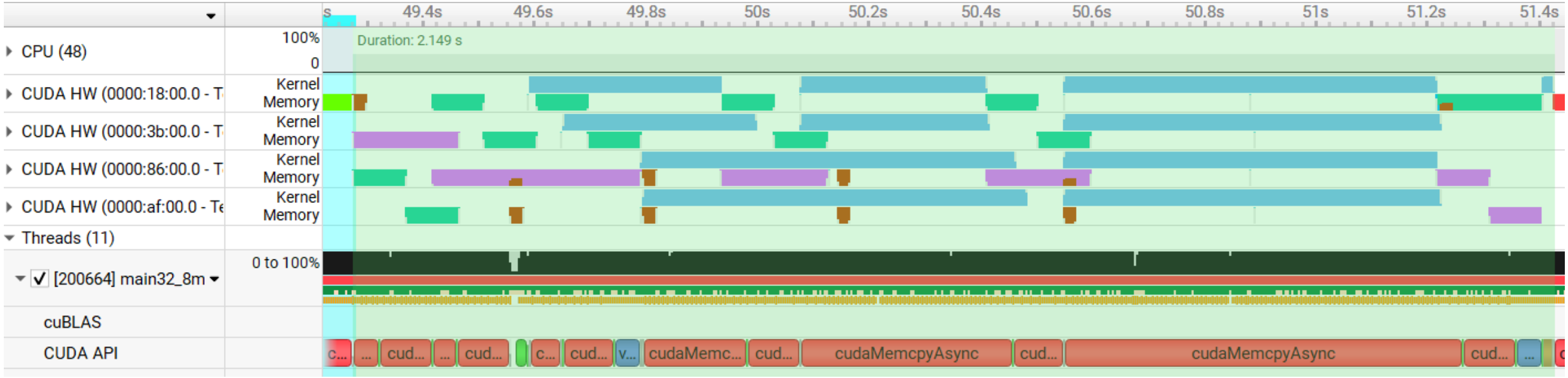
On cHARISMa supercomputer with 4xV100 profiling with Nsight Systems



The profile of the Multi-GPU SGEMM operation on 4 V100 GPUs. Number of elements (N = 65536) in a row (column) of matrices and tile size (Ni = 1024 and 4096). Matrices A, B, and C are stored in devices 1, 2, and 0, respectively.

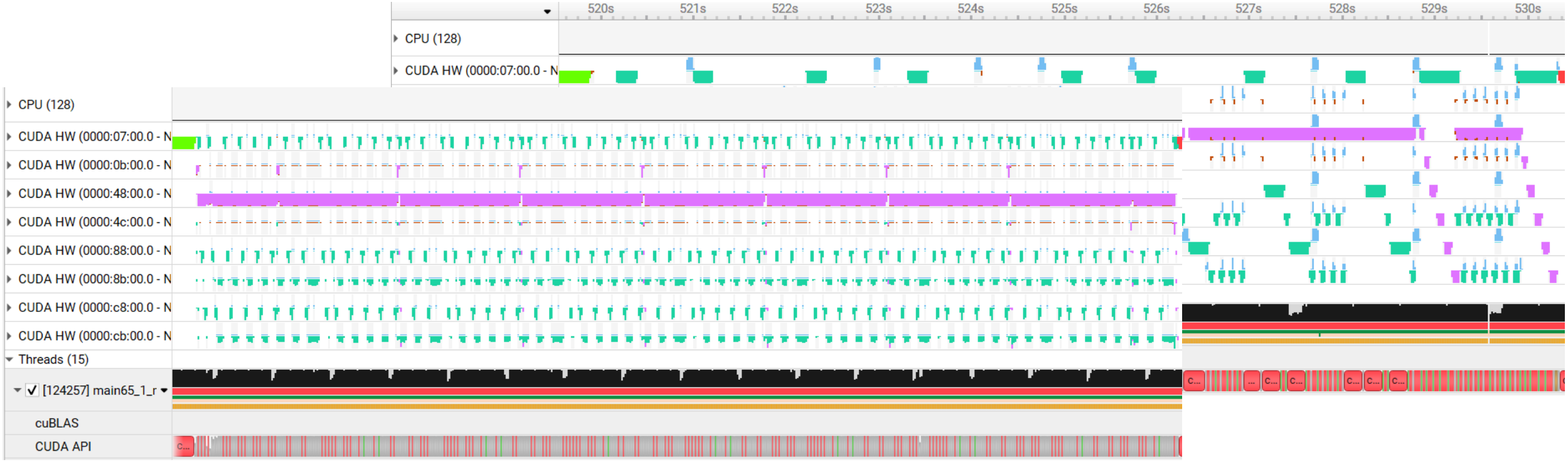
On cHARISMa supercomputer with 4xV100 profiling with Nsight Systems

Synchronous data transfer operations via PCIe



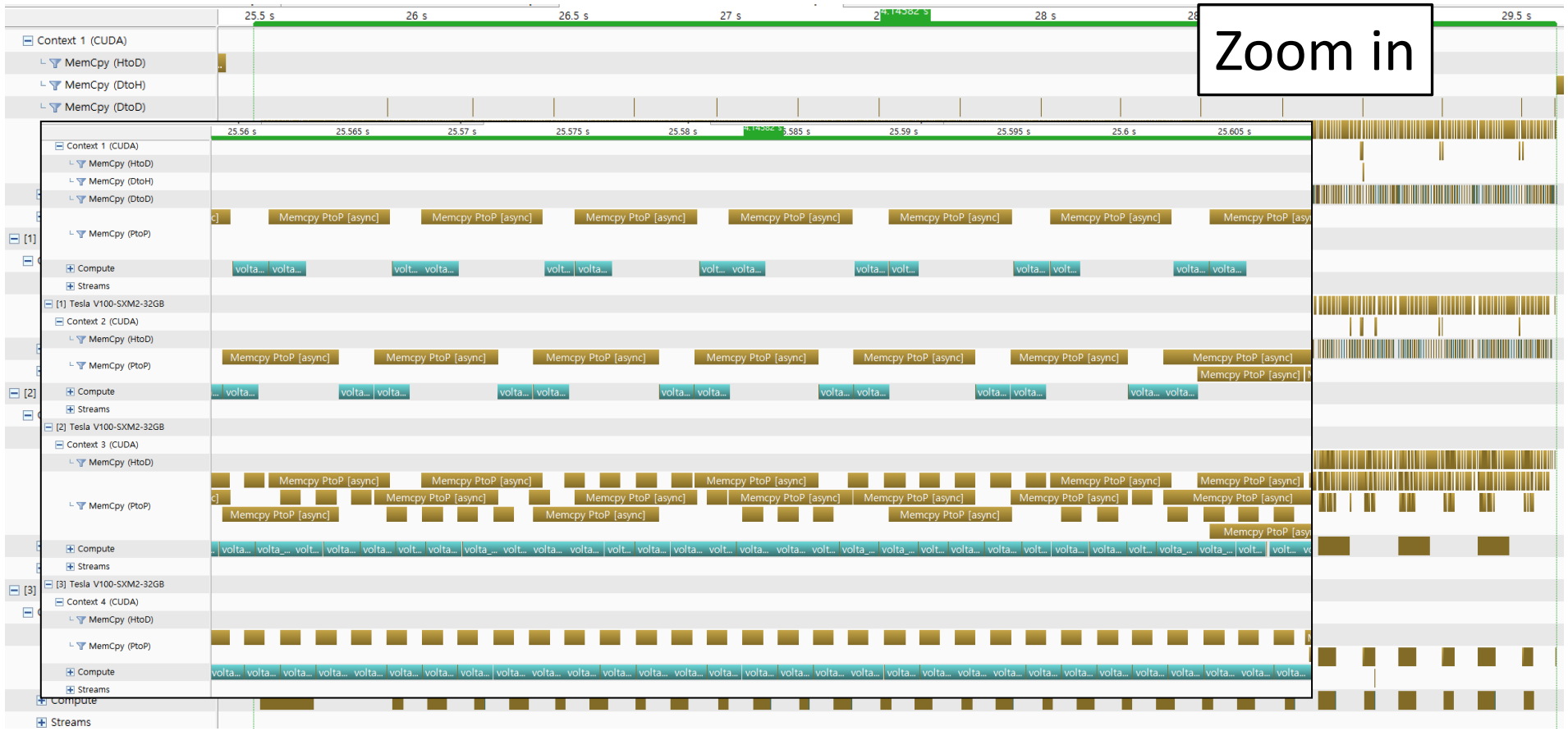
The profile of the Multi-GPU SGEMM operation on 4 V100 GPUs. Number of elements (N = 32768) in a row (column) of matrices and tile size (N_i = 8192). Matrices A, B, and C are stored in devices 1, 2, and 0, respectively.

On cHARISMa supercomputer with 8xA100 profiling with Nsight Systems



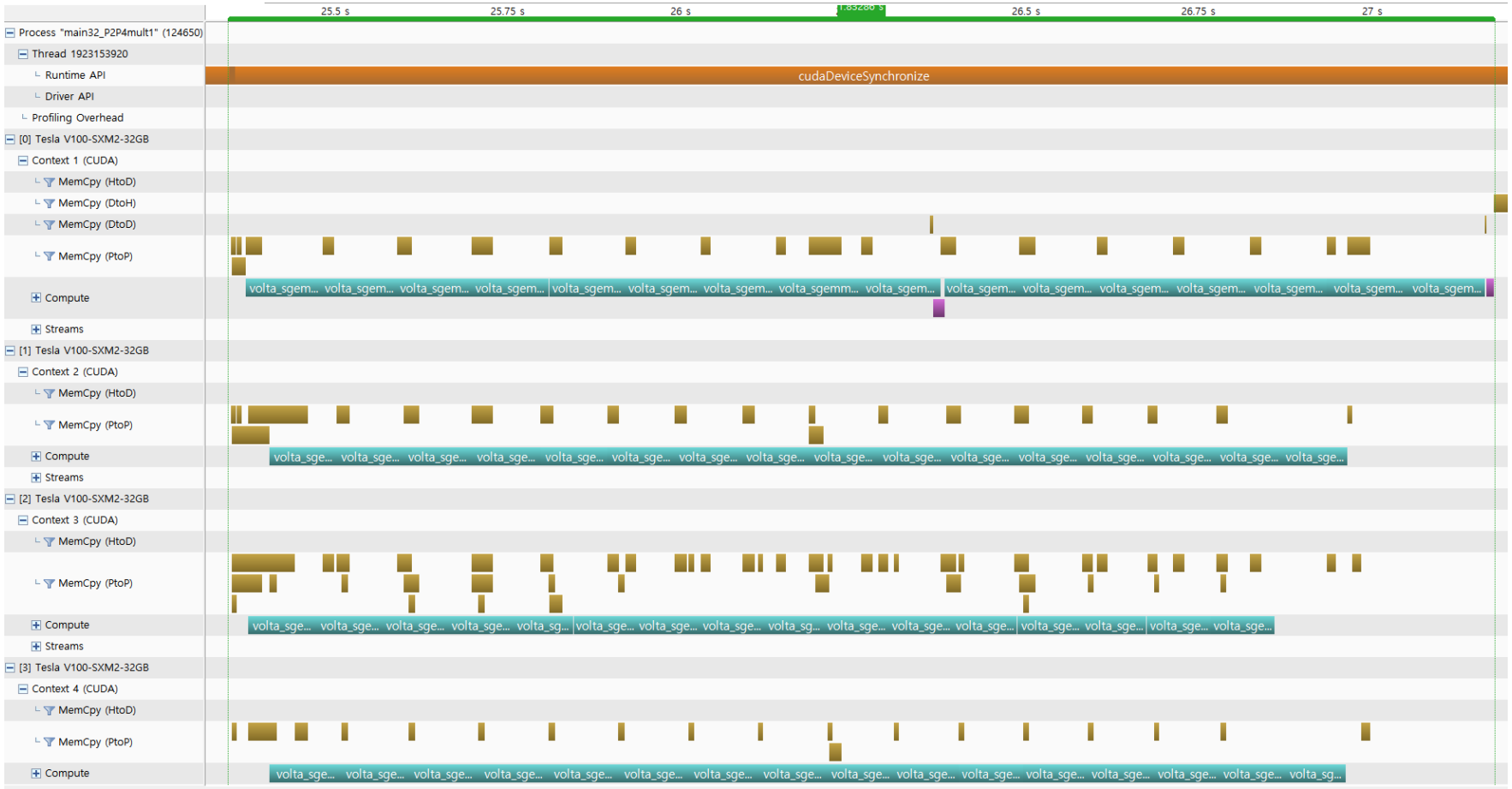
The profile of the Multi-GPU SGEMM operation on 8 A100 GPUs. Number of elements (N = 65536) in a row (column) of matrices and tile size (N_i = 1024 and 8192) and (N_i = 4096). Matrices A, B, and C are stored in devices 1, 2, and 0, respectively.

On IBM Power Systems profiling with Visual Profiler



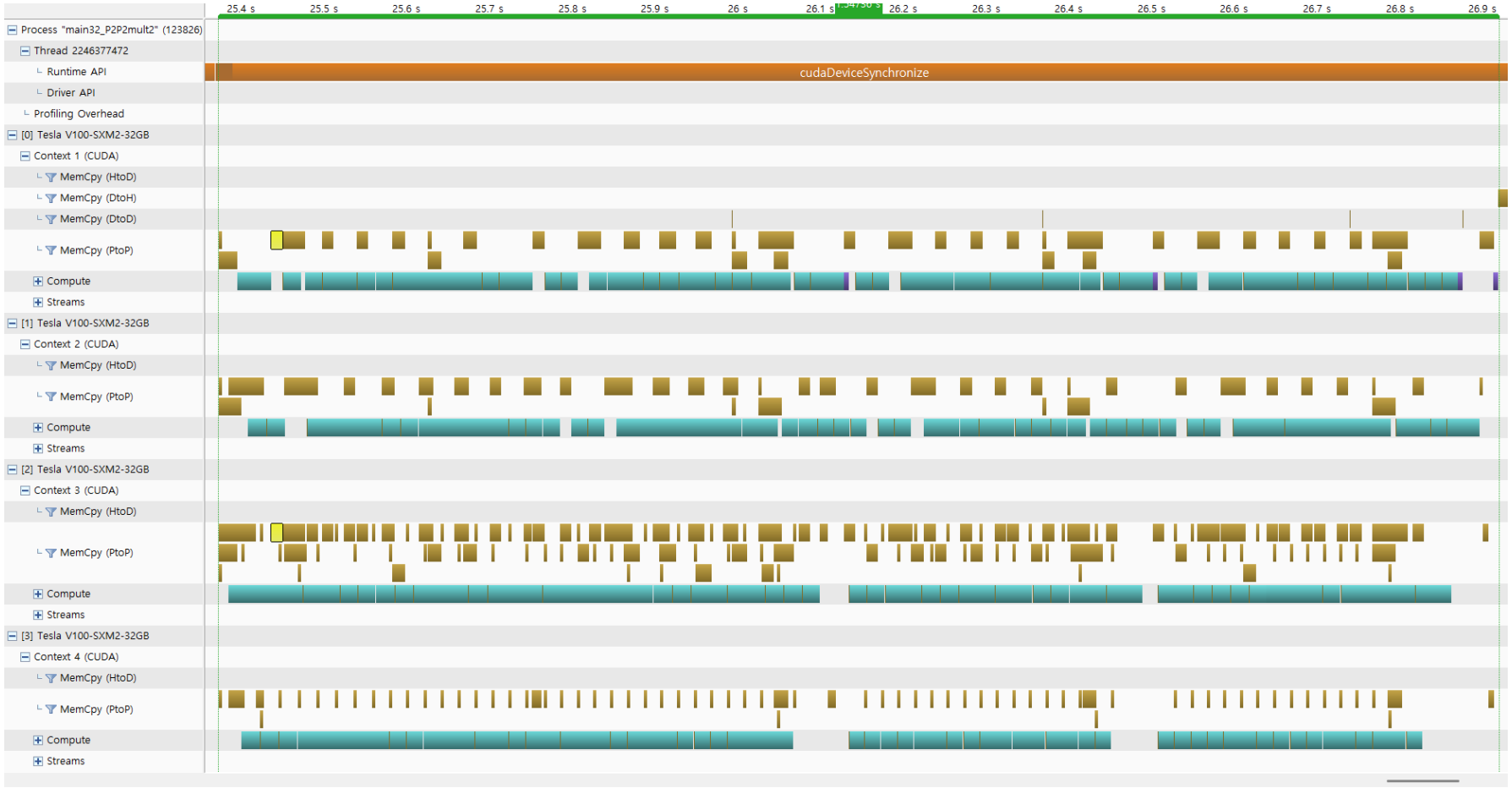
The profile of the Multi-GPU SGEMM operation on IBM POWER 9 server. Number of elements (N = 32768) in a row (column) of matrices, and tile sizes (N_i = 512 and 1024). Matrices A, B, and C are stored in devices 1, 2, and 0, respectively.

On IBM Power Systems profiling with Visual Profiler



The profile of the Multi-GPU SGEMM operation on IBM POWER 9 server. Number of elements (N = 32768) in a row (column) of matrices, and tile sizes (N_i = 4096). Matrices A, B, and C are stored in devices 1, 2, and 0, respectively.

On IBM Power Systems profiling with Visual Profiler



The profile of the Multi-GPU SGEMM operation on IBM POWER 9 server. Number of elements (N = 32768) in a row (column) of matrices, and tile sizes (N_i = 2048 and 4096). Matrices A, B, and C are stored in devices 1, 2, and 0, respectively.

Summary

- Performance peaks: Multi-GPU algorithms on heterogeneous servers can reach up to 76% efficiency, but actual results strongly depend on inter-GPU communication speed.
- Efficiency drops: On cHARISMa's V100 nodes, indirect P2P paths limit performance to about 55% of peak. With 8 A100 GPUs, restrictive tile sizes (due to memory limits) prevent optimal configuration, resulting in substantially degraded performance.
- Key bottlenecks: CPU involvement in D2H/H2D transfers and shared, unidirectional GPU links force synchronous processing, further reducing throughput.
- Additional considerations: For complex workloads, memory bandwidth and the number of asynchronous engines available critically affect sustained performance.