



TDSE Runtime Performance Report

Microsecond-scale execution. Across architectures.

Publication information

TDSE Runtime Performance Report

Document ID	TDSE-RPR-001
Revision	1.0
Publication date	September 23, 2026
Published by	TDSE Technologies

About this publication

This report presents measured TDSE Runtime performance across CPU, GPU, and FPGA platforms. It documents per-step latency, workload scaling, and the execution conditions underlying those results.

Results apply to the software builds, hardware configurations, workloads, and measurement methods identified in this edition. Application-level performance depends on the surrounding integration and operating conditions.

Copyright and attribution

Copyright © 2026 Haikou Shikong Mailuo Technology Co., Ltd. All rights reserved.

TDSE Technologies is a trade name of Haikou Shikong Mailuo Technology Co., Ltd.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form without the prior written permission of Haikou Shikong Mailuo Technology Co., Ltd.

Third-party company and product names are used for identification. They remain the property of their respective owners.

Reference and enquiries

Recommended citation

TDSE Technologies. TDSE Runtime Performance Report. TDSE-RPR-001, revision 1.0, September 23, 2026.

For technical enquiries or requests to review supporting measurement and audit material, contact TDSE Technologies and quote the document ID and revision.

www.tdse.tech · support@tdse.tech

Executive summary

TDSE separates model preparation from repeated time-domain evaluation. The resulting prepared operator combines low per-step cost with parallel execution for larger problems. This report measures that capability across CPU, GPU and FPGA implementations.

4.32 μ s

Low CPU step latency

Lowest measured CPU FP32 median · np=32, nh=1024

25 platforms

Native execution across architectures

CPU, GPU and FPGA · 28 measured configurations

Efficient execution at each time step

Across the tested CPUs, median FP32 step latency is 4.32-17.99 μ s for 32 ports and 1,024 history samples. TDSE evaluates the prepared operator directly on the host, including its input-history contribution. Model preparation precedes the timed steps.

Scalability as operators grow

The larger tested operators show that TDSE can distribute the calculation across multiple GPUs and reduce step latency. At 256 ports and 2,048 history samples, the RTX PRO 6000 Workstation study demonstrates acceleration in both FP32 and FP64, with either device-resident vectors or per-step host transfers. This supports multi-GPU scalability for large operators; smaller workloads often remain faster on one GPU.

The same calculation across architectures

Native TDSE implementations execute on Intel, AMD, Arm and Apple CPUs, NVIDIA CUDA GPUs, and an AMD (Xilinx) FPGA. The measured CPU and GPU matrices cover both single and double precision; FPGA evidence covers three fixed-bitstream FP32 cases. Together, the results demonstrate multiple execution targets for the TDSE calculation, with each platform's scope and timing boundary stated explicitly.

Evidence behind the results. Across 25 platform groups and 28 configurations, 1,299 of 1,299 recorded numerical checks passed. Complete matrices and audit records support the reported measurements.

The following chapters establish the timing method, compare a common reference workload, and examine scaling across larger operators. These prepared-step timings do not establish complete-circuit execution time or real-time deadline compliance.

Contents

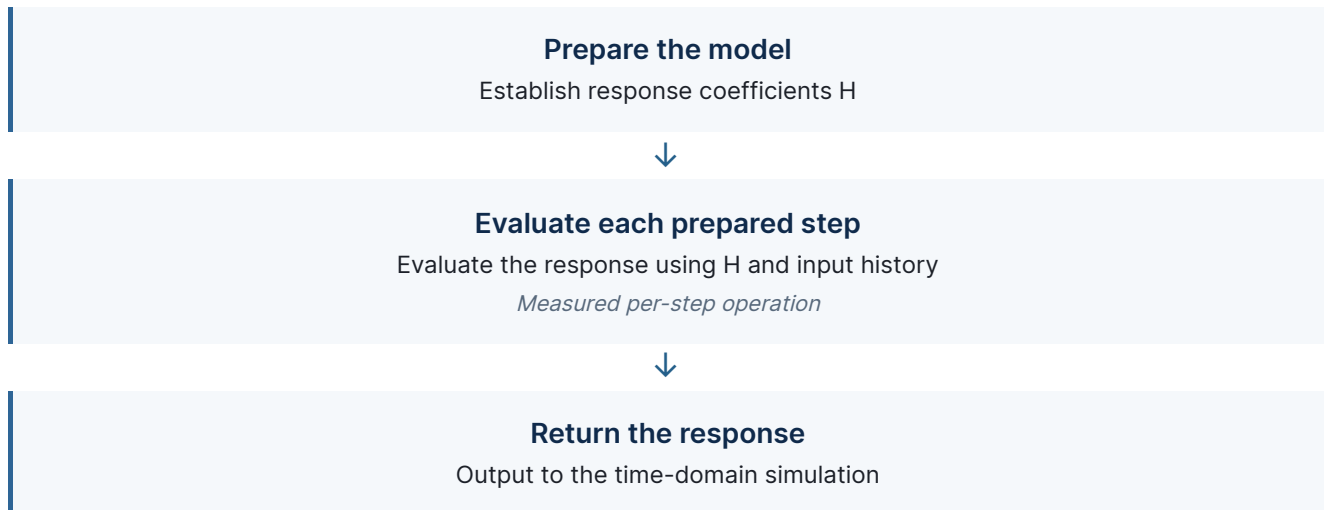
Technical results first. Complete measurements follow for detailed review.

Section	Page
Executive summary	1
TDSE in transient simulation	3
How performance is measured	4
Portability in practice	5
Portable CPU performance	6
GPU and FPGA execution	7
RTX PRO 6000 Workstation scaling analysis	8
Multi-GPU acceleration for large operators	9
Conclusions for transient simulation	10
Measurement notes and data availability	11
Complete CPU and GPU comparison tables	12
Complete platform profiles and result matrices	20

The opening chapters establish TDSE's measured execution cost, parallel capability, and implementation coverage. The final sections retain every platform profile, measured matrix, and variation statistic. 86 PDF pages in total; numbered content runs from 1 to 83.

TDSE in transient simulation

TDSE represents a prepared linear dynamic response through coefficients and input history. At each simulation step, the runtime evaluates that response and returns the model's output to the surrounding simulation.



Preparation separates setup from repeated execution

Model preparation and coefficient loading take place before timing. The prepared operator is then evaluated repeatedly as inputs and their history advance. The report measures this recurring work, including the specified execution boundary for each implementation.

Ports and history define the operator's scale

With np ports and nh history samples, H stores $np \times np \times nh$ coefficients. More ports increase both input and output dimensions; longer histories retain more of the dynamic response. A 256-port, 2,048-sample operator contains 134,217,728 coefficients.

The same calculation across execution targets

CPU, CUDA GPU, and fixed-bitstream FPGA implementations demonstrate different ways to execute the TDSE calculation. The measured timings establish operator-level performance; the surrounding circuit solution, control logic, and I/O remain part of the complete simulation.

How performance is measured

We measure the elapsed time per prepared TDSE step after model preparation and initialization. Each workload specifies the number of ports, the impulse-response history length, and the floating-point precision.

Workload

Ports (np). The number of ports in the TDSE operator.

History (nh). The number of impulse-response history samples. The stored coefficient array H contains $np \times np \times nh$ values; the matrices also give its storage in decimal MB.

Coverage. CPU: np=1–128; GPU: np=8–256. Each uses eight port counts, three history lengths (512, 1024, 2048), and FP32/FP64, totaling 48 cases. FPGA covers three FP32 test cases.

Statistics

Timing repeats. Each repeat times many prepared steps and calculates the arithmetic mean time per step. Where required, additional measurements are processed under the recorded sample-selection rule. Latency is the median of the retained repeat means.

Protocols. CPU/GPU: seven retained repeats, 250 ms target per repeat, 64 warm-up steps. FPGA: five warm-up calls, 30 measured calls, eight steps per call. Different sampling protocols and timed operations limit cross-family comparability.

Variation and numerical checks. CV is the standard deviation of the repeat means divided by their mean. Each measured case is checked against its expected numerical result; the complete matrices report the latency and CV.

Timing boundary

CPU. Host elapsed time around the repeated prepared-step call.

Device-resident GPU. Dynamic vectors remain on the GPU; host elapsed time includes the prepared call and completion wait. This is not kernel-only timing.

Host-resident GPU. Host-observed time includes the required per-step vector transfers. H remains on the GPU in both views.

FPGA. Host-observed time includes the step operation and vector transfers. Loading H and configuring the matching bitstream precede the timed operation.

Reading comparisons

Compare the same np, nh, precision, and timed operations. Results describe the recorded hardware and software configurations; build and driver details accompany each platform.

Using a real-time budget

These are steady-state operator timings. Complete-circuit execution time, tail latency, and deadline compliance require application-level measurement. Further measurement limits are collected in Measurement notes and data availability.

Portability in practice

TDSE is measured across desktop and cloud systems, spanning x86-64 and Arm CPUs, CUDA GPUs, and an FPGA implementation. The table identifies the tested platforms and resources; implementation coverage and qualification boundaries are collected in the report notes.

Measured compute platforms

Platform	Execution target	Resources
CPU platforms		48 cases per configuration
Windows 11 · AMD Ryzen 7 9800X3D	x86-64	8 cores
AWS c8i.2xlarge · 4 cores	x86-64	4 cores
AWS c8i.4xlarge · 8 cores	x86-64	8 cores
AWS c8i.8xlarge · 16 cores	x86-64	16 cores
AWS c8i.metal-48xl · 96 cores	x86-64	96 cores
AWS c8a.2xlarge · 4 cores	x86-64	4 cores
AWS c8a.2xlarge · 8 cores	x86-64	8 cores
AWS c8a.4xlarge · 16 cores	x86-64	16 cores
AWS c8a.metal-24xl · 96 cores	x86-64	96 cores
AWS c7i.4xlarge · 8 cores	x86-64	8 cores
AWS c7a.2xlarge · 8 cores	x86-64	8 cores
AWS c9g.2xlarge · 8 cores	AArch64	8 cores
AWS c9g.4xlarge · 16 cores	AArch64	16 cores
AWS c8g.2xlarge · 8 cores	AArch64	8 cores
AWS c8g.4xlarge · 16 cores	AArch64	16 cores
macOS 27.0 · Apple M5	AArch64	10 cores
GPU platforms		48 cases per configuration
Windows 11 · RTX 5090 (1 GPU)	CUDA CC 12.0	170 SMs
RunPod · RTX 5090 (1 GPU)	CUDA CC 12.0	170 SMs
RunPod · RTX PRO 6000 (1 GPU)	CUDA CC 12.0	188 SMs
RunPod · A100 80GB (1 GPU)	CUDA CC 8.0	108 SMs
RunPod · H100 80GB (1 GPU)	CUDA CC 9.0	132 SMs
RunPod · H200 (1 GPU)	CUDA CC 9.0	132 SMs
RunPod · B200 (1 GPU)	CUDA CC 10.0	148 SMs
RunPod · RTX PRO 6000 Workstation (4 GPUs)	CUDA CC 12.0	4 × 188 SMs
FPGA platform		3 FP32 cases
AWS f2.6xlarge · VU47P (1 FPGA)	FPGA · VU47P	1 FPGA

The 25 platforms contain 28 measured configurations. The formal RTX PRO 6000 Workstation platform includes separate results for selecting 1, 2, 3, or 4 GPUs. Each CPU and GPU configuration has 48 cases, with different port-count grids; the FPGA covers three FP32 cases. The next two pages present a common 32-port reference workload.

Portable CPU performance

Low TDSE step latency is not confined to dedicated accelerators. For a 32-port operator with 1,024 history samples, the desktop and cloud CPUs below execute the prepared step in microseconds, in both single and double precision.

Platform	Execution target	Resources	FP32 (μ s)	FP64 (μ s)
Windows 11 · AMD Ryzen 7 9800X3D	x86-64	8 cores	5.68	11.86
AWS c8i.2xlarge · 4 cores	x86-64	4 cores	10.70	75.19
AWS c8i.4xlarge · 8 cores	x86-64	8 cores	6.56	11.35
AWS c8i.8xlarge · 16 cores	x86-64	16 cores	7.78	12.06
AWS c8i.metal-48xl · 96 cores	x86-64	96 cores	13.29	15.90
AWS c8a.2xlarge · 4 cores	x86-64	4 cores	11.59	29.26
AWS c8a.2xlarge · 8 cores	x86-64	8 cores	4.32	14.83
AWS c8a.4xlarge · 16 cores	x86-64	16 cores	6.48	11.32
AWS c8a.metal-24xl · 96 cores	x86-64	96 cores	11.11	14.25
AWS c7i.4xlarge · 8 cores	x86-64	8 cores	7.26	13.26
AWS c7a.2xlarge · 8 cores	x86-64	8 cores	7.04	16.63
AWS c9g.2xlarge · 8 cores	AArch64	8 cores	13.33	24.67
AWS c9g.4xlarge · 16 cores	AArch64	16 cores	11.82	21.87
AWS c8g.2xlarge · 8 cores	AArch64	8 cores	17.99	31.91
AWS c8g.4xlarge · 16 cores	AArch64	16 cores	13.67	25.55
macOS 27.0 · Apple M5	AArch64	10 cores	5.07	10.11

prepared-step median at $np=32$, $nh=1024$. The complete platform matrices also report CV. The c8i.2xlarge FP64 value is retained as measured; its repeat range, thread configuration and unresolved performance discontinuity are discussed below.

Reading the CPU results

c8i.2xlarge. FP64 at $np=32$, $nh=1024$ has median 75.19 μ s (repeat means 74.95–76.41 μ s; CV 0.61%). The retained repeats confirm the high latency within this run. The packet records 4 configured BLAS workers and 4 dispatch tasks at this shape; SMT is enabled. This is not a controlled core-count scaling experiment. Cache, dispatch and scheduling contributions were not isolated.

Apple M5. FP32 at $nh=1024$ falls from 4.46 μ s at $np=8$ to 1.80 μ s at $np=16$ (CV 0.11% and 0.18%). Latency is therefore not monotonic in operator size. Accelerate manages internal threading and macOS places workers; this capture does not identify the kernel dispatch or P/E-core placement responsible for the change.

GPU and FPGA execution

GPU and FPGA implementations extend TDSE beyond host-CPU execution. The table uses the same 32-port, 1,024-sample reference workload. The larger-workload results that follow show where distributing the calculation across GPUs begins to help.

Platform	Execution target	Resources	FP32 (μs)	FP64 (μs)
GPU platforms				
Windows 11 · RTX 5090 (1 GPU)	CUDA CC 12.0	170 SMs	20.25	22.07
RunPod · RTX 5090 (1 GPU)	CUDA CC 12.0	170 SMs	14.46	17.28
RunPod · RTX PRO 6000 (1 GPU)	CUDA CC 12.0	188 SMs	20.03	23.44
RunPod · A100 80GB (1 GPU)	CUDA CC 8.0	108 SMs	27.66	27.80
RunPod · H100 80GB (1 GPU)	CUDA CC 9.0	132 SMs	21.52	21.53
RunPod · H200 (1 GPU)	CUDA CC 9.0	132 SMs	20.90	20.24
RunPod · B200 (1 GPU)	CUDA CC 10.0	148 SMs	23.85	21.94
RunPod · RTX PRO 6000 Workstation (4 GPUs)	CUDA CC 12.0	4 × 188 SMs	57.03	59.12
FPGA platforms				
AWS f2.6xlarge · VU47P (1 FPGA)	FPGA · VU47P	1 FPGA	12.49	N/A

GPU rows use device-resident medians; the FPGA row uses its fixed-bitstream host-observed boundary at np=32, nh=1024. The different timed operations preclude a cross-family ranking. The RTX PRO 6000 Workstation × 4 row times the complete four-GPU configuration. At this small workload, multi-GPU overhead outweighs the benefit; see the following scaling study for larger operators.

Larger operators reveal the parallel benefit

The reference workload above is small enough that adding GPUs need not reduce step latency. More ports and a longer history provide more computation to distribute. The scaling study therefore examines the larger workload.

At 256 ports and 2,048 history samples, the formal multi-GPU measurements show speedups of 4.19× with 4 GPUs in FP32 and 2.93× with 4 GPUs in FP64, relative to the one-GPU configuration in this scaling study. These are the lowest-latency selections among the tested GPU counts, using device-resident timing.

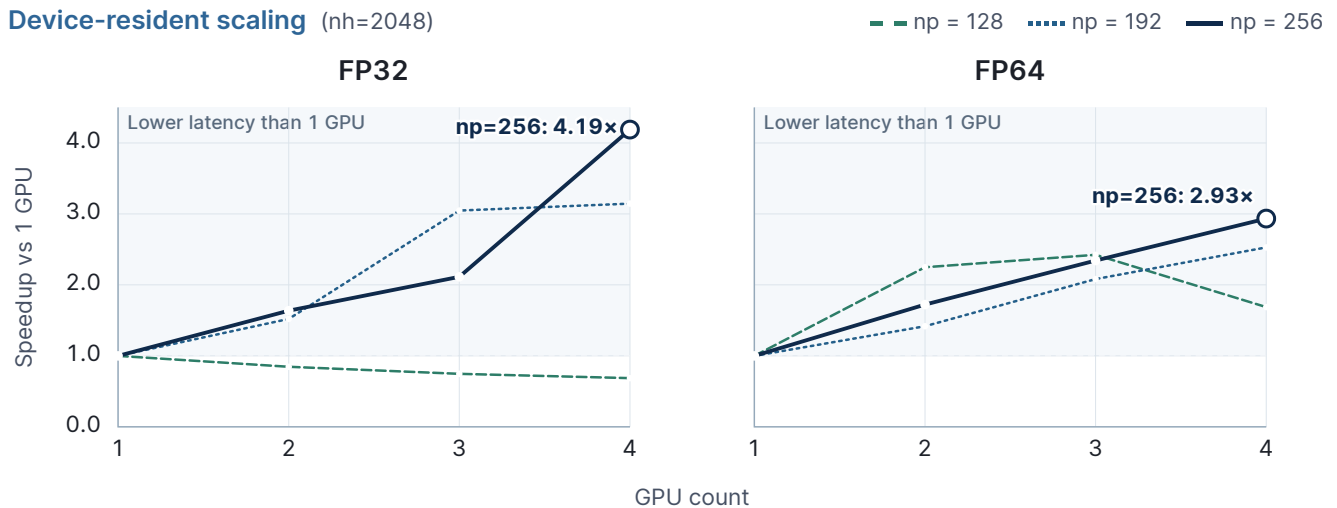
FPGA results describe a fixed implementation

The AMD (Xilinx) FPGA results cover three FP32 configurations, each with a matching bitstream. The value above includes per-step vector transfers. It demonstrates execution of the TDSE calculation on FPGA hardware; the CPU/GPU matrices cover a broader range of workloads and both precisions.

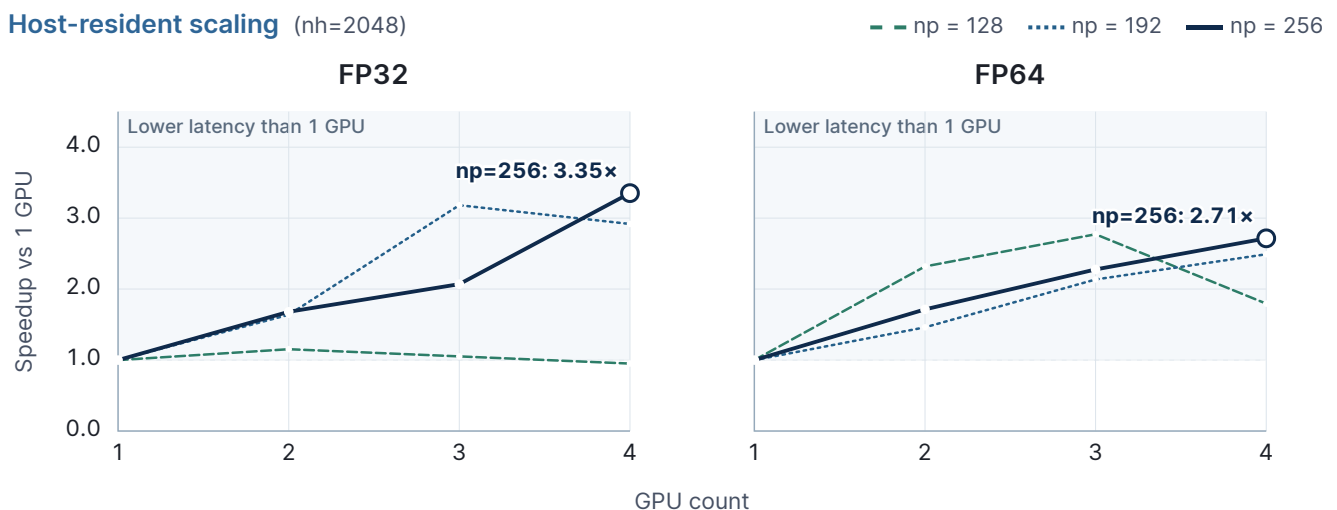
RTX PRO 6000 Workstation scaling analysis

TDSE distributes larger prepared operators across multiple GPUs, with measured acceleration in both device- and host-resident execution. The RTX PRO 6000 Workstation curves cover np=128, 192, and 256 at nh=2048.

Device-resident scaling (nh=2048)



Host-resident scaling (nh=2048)



Parallel acceleration in both residency modes

At 256 ports and 2,048 history samples, TDSE accelerates both precisions even when per-step host transfers are included. FP32 achieves 4.19× speedup with device-resident vectors (4 GPUs) and 3.35× with host transfers (4 GPUs). The smaller workloads show why scaling depends on operator size: additional GPUs do not consistently reduce latency at 128 and 192 ports.

Speedup is the one-GPU median divided by the corresponding multi-GPU median for the same shape, precision, and residency mode. The following page places these larger operators in the context of the full matrix.

Multi-GPU acceleration for large operators

A 256-port operator with 2,048 history samples contains 134,217,728 coefficients. TDSE evaluates this operator in tens to hundreds of microseconds in the measured RTX PRO 6000 Workstation configurations.

From one GPU to four

Device-resident median step latency at np=256, nh=2048:

Precision	1 GPU	2 GPUs	3 GPUs	4 GPUs
FP32	347.65 μ s	212.43 μ s	164.47 μ s	82.99 μ s
FP64	679.73 μ s	394.53 μ s	290.11 μ s	231.63 μ s

The operator contains 536.87 MB of FP32 coefficients or 1,073.74 MB in FP64. Preparation and coefficient loading precede timing; these results quantify the recurring evaluation of the prepared operator.

Scaling across the complete workload matrix

GPUs	Median speedup across test cases	Cases faster than 1 GPU
1	1.000×	Baseline
2	0.654×	12/48
3	0.411×	12/48
4	0.373×	12/48

Device-resident results across all 48 test cases, with equal weight per case. Speedup is computed against the matching one-GPU case before taking the median. This aggregate does not represent an application workload mix.

The same cases improve at 2, 3 and 4 GPUs: FP32 np=192, nh=1024/2048; FP32 np=256, nh=1024/2048; FP64 np=96, nh=2048; FP64 np=128, nh=2048; FP64 np=192, nh=512/1024/2048; FP64 np=256, nh=512/1024/2048.

Where parallel execution helps

The larger-operator results demonstrate that TDSE can convert parallel GPU execution into lower step latency. The complete matrix also shows the boundary: most smaller workloads remain faster on one GPU. The large-workload speedups and the full-matrix median therefore describe different parts of the same experiment.

Both residency views show that the benefit extends beyond device-resident vectors to execution that includes per-step host transfers.

All four GPU-count configurations include complete 48-row matrices and full audit data. Access details appear in the report notes.

Conclusions for transient simulation

The measurements demonstrate three capabilities of TDSE: microsecond-scale prepared-step execution, parallel acceleration of larger operators, and native implementations across CPU, GPU, and FPGA hardware.

Microsecond-scale evaluation on CPUs

At 32 ports and 1,024 history samples, measured CPU medians span 4.32-75.19 μ s across FP32 and FP64. TDSE evaluates the prepared operator directly on the host, including its input-history contribution, without requiring an accelerator.

The full matrices extend the evidence across port counts, history lengths, and precisions. They show how per-step latency changes beyond a single reference case.

Larger operators benefit from parallel execution

The RTX PRO 6000 Workstation measurements show lower latency for the 256-port, 2,048-sample operator as execution expands from one to four GPUs. Both FP32 and FP64 benefit, including host-resident execution with per-step vector transfers. This extends the demonstrated TDSE capability to larger operators; the gain varies across the tested workloads.

One technology across computing architectures

Native CPU measurements span Windows, Linux, and macOS on Apple Silicon. CUDA measurements span Windows and Linux, while the FPGA results demonstrate a fixed-bitstream implementation of the TDSE calculation.

The FPGA records 12.49 μ s in FP32 at the 32-port reference workload, including per-step vector transfers. Its three test cases demonstrate another hardware implementation to the broader CPU/GPU evidence.

Low per-step latency for time-domain simulation. TDSE combines fast prepared-operator evaluation with demonstrated parallel and cross-platform execution. These measurements establish the operator-level performance foundation; complete-circuit timing and real-time deadlines remain application-level properties.

Measurement notes and data availability

These notes explain how to use the reported timings and obtain the supporting data. Measurement definitions are given on page 4; each platform section identifies its recorded configuration and results.

Measurement scope

Results measure steady-state prepared steps after model preparation, coefficient loading, and accelerator configuration. Host-resident GPU and FPGA timings include the stated per-step vector transfers.

Latency and CV summarize retained repeat means under each run's recorded sampling policy. Tail latency, deadline-miss rates, dynamic coefficient updates, batching, power, and operating cost were not characterized. This report evaluates TDSE execution capability rather than procurement cost efficiency.

H alone is not peak memory use: history, workspace and runtime allocations also consume capacity; untested memory sizes are not qualified.

Implementation coverage

Results apply to the recorded SDK builds, operating systems, drivers, and hardware configurations. A common report revision does not imply identical binaries or timing boundaries across platforms.

FPGA coverage comprises three FP32 test cases using fixed bitstreams. The native macOS developer build uses development authorization; these measurements do not qualify a signed and notarized customer package or grant a license.

Comparison and verification

Comparisons use matching port counts, history lengths, precisions, and timed operations. The RTX PRO 6000 Workstation study uses its one-GPU configuration as the reference for the 1/2/3/4-GPU results.

Numerical pass counts refer to checks against each measured case's reference result and tolerance. They establish the reported operator checks, not complete-circuit accuracy or real-time deadline compliance.

Evidence and data access

The supporting data contain the measured matrices, experiment inventory, captured configurations, checksums, and audit records. Hardware topology is recorded as reported by the operating system; variation across instances was not characterized. Revision remains 1.0; editions and corrections are distinguished by publication date and the PDF hash in the build receipt.

Request the supporting materials using document ID TDSE-RPR-001, Revision 1.0, from support@tdse.tech. Measurement definitions appear on page 4; detailed results accompany each platform.

CPU performance comparison

FP32 · 512 history samples · Host-elapsed timing

Lower latency  Higher latency

Shared log color scale: 0.01-10,000 μs

Median step latency (μs) · Lower is faster

Fixed platform order across all history lengths and precisions

Platform	np=1	np=2	np=4	np=8	np=16	np=32	np=64	np=128
9800X3D · 8 cores	0.14	0.47	1.00	2.21	2.99	3.98	11.45	55.58
c8i.2xlarge · 4 cores	0.09	1.32	1.43	1.90	2.45	6.03	45.23	486.38
c8i.4xlarge · 8 cores	0.11	1.24	1.43	1.85	2.80	4.13	11.43	248.12
c8i.8xlarge · 16 cores	0.09	1.11	1.46	1.94	3.86	5.24	7.74	76.00
c8i.metal-48xl · 96 cores	0.09	1.96	2.14	2.77	5.01	9.66	18.42	33.45
c8a.2xlarge · 4 cores	0.06	0.40	0.57	0.89	1.38	4.25	26.23	116.69
c8a.2xlarge · 8 cores	0.07	0.41	0.55	0.82	1.48	2.45	11.01	50.93
c8a.4xlarge · 16 cores	0.07	1.35	1.67	2.07	2.63	4.14	6.39	28.91
c8a.metal-24xl · 96 cores	0.07	1.50	2.03	2.98	4.53	7.43	13.26	23.24
c7i.4xlarge · 8 cores	0.10	1.55	2.17	2.39	3.43	4.85	12.54	163.55
c7a.2xlarge · 8 cores	0.11	0.67	0.93	1.44	2.73	3.97	15.15	67.20
c9g.2xlarge · 8 cores	0.10	0.40	0.74	1.67	3.63	7.08	24.42	148.80
c9g.4xlarge · 16 cores	0.10	0.42	0.76	1.62	3.56	6.80	13.72	72.35
c8g.2xlarge · 8 cores	0.14	0.49	0.89	1.95	4.09	9.41	31.85	185.15
c8g.4xlarge · 16 cores	0.14	0.52	0.92	2.02	4.17	7.80	17.96	113.74
Apple M5 · 10 cores	0.06	0.51	0.93	2.17	0.98	2.63	9.61	420.50

Bold values mark the lowest displayed median in each column, including ties. Platform labels link to their full profiles and thread configuration. CPU observations accompany the reference comparison.

Step latency across operator sizes

The shortest tested history highlights how step latency changes as the number of ports grows. At 128 ports and 512 history samples, TDSE evaluates an operator with 8,388,608 FP32 coefficients. Each row shows how one platform's measured latency changes across these workloads.

CPU performance comparison

FP32 · 1,024 history samples · Host-elapsed timing

Lower latency Higher latency

Shared log color scale: 0.01-10,000 μ s

Median step latency (μ s) · Lower is faster

Fixed platform order across all history lengths and precisions

Platform	np=1	np=2	np=4	np=8	np=16	np=32	np=64	np=128
9800X3D · 8 cores	0.16	0.54	1.09	2.52	3.89	5.68	26.47	107.99
c8i.2xlarge · 4 cores	0.10	1.31	1.55	2.52	3.67	10.70	243.40	1,014.45
c8i.4xlarge · 8 cores	0.12	1.29	1.63	2.47	4.10	6.56	126.33	498.25
c8i.8xlarge · 16 cores	0.10	1.19	1.60	2.54	5.46	7.78	14.50	258.90
c8i.metal-48xl · 96 cores	0.10	2.12	2.41	3.49	7.76	13.29	29.66	55.28
c8a.2xlarge · 4 cores	0.08	0.46	0.68	1.37	2.31	11.59	42.71	771.68
c8a.2xlarge · 8 cores	0.08	0.49	0.67	1.03	2.43	4.32	25.27	748.91
c8a.4xlarge · 16 cores	0.08	1.42	1.81	2.35	3.35	6.48	14.37	67.56
c8a.metal-24xl · 96 cores	0.08	1.64	2.16	3.83	6.30	11.11	19.70	43.01
c7i.4xlarge · 8 cores	0.11	1.61	2.30	3.14	4.75	7.26	83.48	327.36
c7a.2xlarge · 8 cores	0.13	0.77	1.16	1.90	4.48	7.04	30.80	1,002.47
c9g.2xlarge · 8 cores	0.14	0.72	1.40	2.32	6.03	13.33	69.00	362.01
c9g.4xlarge · 16 cores	0.14	0.74	1.42	2.24	4.29	11.82	25.05	432.31
c8g.2xlarge · 8 cores	0.18	0.87	1.67	2.57	6.99	17.99	95.21	423.94
c8g.4xlarge · 16 cores	0.18	0.89	1.68	2.60	4.97	13.67	34.26	386.72
Apple M5 · 10 cores	0.08	0.97	1.82	4.46	1.80	5.07	137.79	873.13

Bold values mark the lowest displayed median in each column, including ties. Platform labels link to their full profiles and thread configuration. CPU observations accompany the reference comparison.

Step latency across operator sizes

This history length connects the 32-port reference workload to the larger operators in the full matrix. At 128 ports and 1,024 history samples, TDSE evaluates an operator with 16,777,216 FP32 coefficients. Each row shows how one platform's measured latency changes across these workloads.

CPU performance comparison

FP32 · 2,048 history samples · Host-elapsed timing

Lower latency Higher latency

Shared log color scale: 0.01-10,000 μ s

Median step latency (μ s) · Lower is faster

Fixed platform order across all history lengths and precisions

Platform	np=1	np=2	np=4	np=8	np=16	np=32	np=64	np=128
9800X3D · 8 cores	0.22	0.65	1.38	3.09	5.72	11.85	53.81	1,004.55
c8i.2xlarge · 4 cores	0.13	1.46	2.15	3.78	6.12	71.89	499.04	1,861.05
c8i.4xlarge · 8 cores	0.15	1.44	2.19	3.34	6.59	11.37	246.40	1,010.43
c8i.8xlarge · 16 cores	0.13	1.35	2.23	3.45	9.06	12.78	131.48	521.60
c8i.metal-48xl · 96 cores	0.13	1.95	3.02	4.80	14.41	22.52	52.52	181.20
c8a.2xlarge · 4 cores	0.10	0.64	0.92	2.31	4.19	28.64	114.60	2,059.03
c8a.2xlarge · 8 cores	0.10	0.65	0.92	1.70	4.36	11.78	51.71	2,101.55
c8a.4xlarge · 16 cores	0.10	1.57	2.10	3.12	5.10	11.17	29.35	785.77
c8a.metal-24xl · 96 cores	0.10	1.69	2.72	5.08	9.85	18.39	35.77	90.24
c7i.4xlarge · 8 cores	0.14	1.84	2.94	4.00	7.52	12.63	164.50	824.50
c7a.2xlarge · 8 cores	0.18	1.01	1.60	2.78	7.87	16.01	69.62	2,901.14
c9g.2xlarge · 8 cores	0.22	1.35	2.02	3.61	11.12	24.28	178.43	863.20
c9g.4xlarge · 16 cores	0.22	1.37	1.94	3.51	6.76	21.82	194.16	1,108.26
c8g.2xlarge · 8 cores	0.27	1.60	2.18	4.16	12.98	32.49	214.71	920.74
c8g.4xlarge · 16 cores	0.27	1.63	2.28	4.06	7.70	25.50	153.07	737.54
Apple M5 · 10 cores	0.13	1.90	3.64	8.68	3.42	9.98	606.07	2,251.36

Bold values mark the lowest displayed median in each column, including ties. Platform labels link to their full profiles and thread configuration. CPU observations accompany the reference comparison.

Step latency across operator sizes

The longest tested history shows the cost of evaluating the largest coefficient arrays in the CPU study. At 128 ports and 2,048 history samples, TDSE evaluates an operator with 33,554,432 FP32 coefficients. Each row shows how one platform's measured latency changes across these workloads.

CPU performance comparison

FP64 · 512 history samples · Host-elapsed timing

Lower latency Higher latency

Shared log color scale: 0.01-10,000 μ s

Median step latency (μ s) · Lower is faster

Fixed platform order across all history lengths and precisions

Platform	np=1	np=2	np=4	np=8	np=16	np=32	np=64	np=128
9800X3D · 8 cores	0.07	0.52	1.09	2.48	3.82	5.99	25.64	115.72
c8i.2xlarge · 4 cores	0.09	1.38	1.67	2.44	3.88	10.52	240.92	972.84
c8i.4xlarge · 8 cores	0.11	1.24	1.59	2.42	4.14	6.46	71.64	484.47
c8i.8xlarge · 16 cores	0.09	1.00	1.52	2.30	4.14	7.59	13.70	256.52
c8i.metal-48xl · 96 cores	0.09	1.70	2.51	3.22	5.51	10.09	19.71	38.83
c8a.2xlarge · 4 cores	0.08	0.45	0.65	1.33	2.29	10.75	47.76	752.98
c8a.2xlarge · 8 cores	0.08	0.47	0.64	1.18	2.43	4.26	32.57	743.91
c8a.4xlarge · 16 cores	0.08	1.39	1.56	2.47	3.38	6.68	14.49	56.70
c8a.metal-24xl · 96 cores	0.08	1.61	1.99	3.61	5.38	9.80	16.04	36.11
c7i.4xlarge · 8 cores	0.10	1.68	2.04	2.67	4.52	7.04	83.38	318.61
c7a.2xlarge · 8 cores	0.12	0.76	1.15	1.83	4.38	6.95	33.05	1,042.15
c9g.2xlarge · 8 cores	0.14	0.72	1.41	2.32	6.12	12.91	49.13	379.52
c9g.4xlarge · 16 cores	0.14	0.73	1.52	2.35	6.09	11.87	24.76	392.06
c8g.2xlarge · 8 cores	0.18	0.86	1.67	2.58	6.97	16.98	94.97	409.49
c8g.4xlarge · 16 cores	0.18	0.88	1.68	2.58	7.03	13.73	56.28	379.26
Apple M5 · 10 cores	0.08	0.72	1.34	0.60	1.45	5.21	121.45	837.82

Bold values mark the lowest displayed median in each column, including ties. Platform labels link to their full profiles and thread configuration. CPU observations accompany the reference comparison.

Step latency across operator sizes

The shortest tested history highlights how step latency changes as the number of ports grows. At 128 ports and 512 history samples, TDSE evaluates an operator with 8,388,608 FP64 coefficients. Each row shows how one platform's measured latency changes across these workloads.

CPU performance comparison

FP64 · 1,024 history samples · Host-elapsed timing

Lower latency Higher latency

Shared log color scale: 0.01-10,000 μs

Median step latency (μs) · Lower is faster

Fixed platform order across all history lengths and precisions

Platform	np=1	np=2	np=4	np=8	np=16	np=32	np=64	np=128
9800X3D · 8 cores	0.09	0.63	1.40	3.08	5.54	11.86	56.35	968.28
c8i.2xlarge · 4 cores	0.12	1.50	2.24	3.75	6.14	75.19	483.77	1,746.78
c8i.4xlarge · 8 cores	0.14	1.40	2.13	3.31	6.45	11.35	258.23	945.08
c8i.8xlarge · 16 cores	0.12	1.14	2.02	3.25	6.30	12.06	112.48	513.31
c8i.metal-48xl · 96 cores	0.12	1.84	2.78	4.23	8.02	15.90	30.89	176.58
c8a.2xlarge · 4 cores	0.10	0.56	0.90	2.27	4.15	29.26	113.36	2,030.65
c8a.2xlarge · 8 cores	0.10	0.60	1.06	1.53	4.33	14.83	71.47	2,077.68
c8a.4xlarge · 16 cores	0.10	1.56	2.17	3.16	4.90	11.32	29.52	763.78
c8a.metal-24xl · 96 cores	0.11	1.64	2.52	4.69	8.43	14.25	29.75	89.00
c7i.4xlarge · 8 cores	0.13	1.83	2.64	3.77	7.16	13.26	181.36	896.72
c7a.2xlarge · 8 cores	0.17	1.00	1.59	2.71	7.90	16.63	167.25	3,062.74
c9g.2xlarge · 8 cores	0.21	1.34	2.95	3.57	11.14	24.67	159.48	874.95
c9g.4xlarge · 16 cores	0.21	1.36	2.96	3.54	6.91	21.87	114.45	1,032.21
c8g.2xlarge · 8 cores	0.27	1.60	3.57	4.02	12.86	31.91	183.80	919.09
c8g.4xlarge · 16 cores	0.27	1.62	3.60	4.06	7.81	25.55	98.31	705.11
Apple M5 · 10 cores	0.12	1.39	2.66	1.02	2.73	10.11	425.32	2,787.57

Bold values mark the lowest displayed median in each column, including ties. Platform labels link to their full profiles and thread configuration. CPU observations accompany the reference comparison.

Step latency across operator sizes

This history length connects the 32-port reference workload to the larger operators in the full matrix. At 128 ports and 1,024 history samples, TDSE evaluates an operator with 16,777,216 FP64 coefficients. Each row shows how one platform's measured latency changes across these workloads.

CPU performance comparison

FP64 · 2,048 history samples · Host-elapsed timing

Lower latency  Higher latency

Shared log color scale: 0.01-10,000 μs

Median step latency (μs) · Lower is faster

Fixed platform order across all history lengths and precisions

Platform	np=1	np=2	np=4	np=8	np=16	np=32	np=64	np=128
9800X3D · 8 cores	0.13	0.89	1.85	4.00	9.45	25.51	111.77	3,841.08
c8i.2xlarge · 4 cores	0.18	2.08	3.02	6.28	11.17	233.28	997.39	3,811.16
c8i.4xlarge · 8 cores	0.19	1.97	3.05	5.05	11.39	129.22	490.55	1,994.11
c8i.8xlarge · 16 cores	0.18	1.72	2.99	4.92	9.55	26.36	277.30	1,014.00
c8i.metal-48xl · 96 cores	0.17	2.47	3.79	6.21	14.01	26.12	110.55	452.07
c8a.2xlarge · 4 cores	0.14	0.83	1.40	4.15	12.83	45.95	795.67	4,562.18
c8a.2xlarge · 8 cores	0.14	0.85	1.41	2.78	8.11	31.39	737.98	4,701.03
c8a.4xlarge · 16 cores	0.14	1.87	2.79	4.60	7.80	20.74	329.98	2,129.87
c8a.metal-24xl · 96 cores	0.15	1.99	3.06	6.66	12.82	28.57	61.13	292.61
c7i.4xlarge · 8 cores	0.19	2.47	3.78	5.70	12.40	87.97	336.09	2,328.40
c7a.2xlarge · 8 cores	0.30	1.44	2.48	4.48	14.97	34.41	1,041.99	6,928.73
c9g.2xlarge · 8 cores	0.37	2.60	3.25	6.07	21.17	77.92	420.00	1,743.84
c9g.4xlarge · 16 cores	0.37	2.62	3.17	6.08	11.84	41.94	437.67	2,567.15
c8g.2xlarge · 8 cores	0.45	3.08	3.72	7.04	24.74	96.82	450.82	1,930.25
c8g.4xlarge · 16 cores	0.45	3.10	3.74	7.01	13.68	64.54	344.18	1,438.69
Apple M5 · 10 cores	0.13	2.74	5.25	1.85	5.23	148.93	1,590.99	6,073.46

Bold values mark the lowest displayed median in each column, including ties. Platform labels link to their full profiles and thread configuration. CPU observations accompany the reference comparison.

Step latency across operator sizes

The longest tested history shows the cost of evaluating the largest coefficient arrays in the CPU study. At 128 ports and 2,048 history samples, TDSE evaluates an operator with 33,554,432 FP64 coefficients. Each row shows how one platform's measured latency changes across these workloads.

GPU performance comparison

FP32 · Device-resident median step latency (μs) · Lower is faster

Lower latencyHigher latency

Shared log color scale: 10.00-1,000 μs

512 history samples (nh)

Fixed platform order across all history lengths and precisions

Platform	np=8	np=16	np=32	np=64	np=96	np=128	np=192	np=256
RTX 5090 (Windows)	20.35	21.98	21.59	19.97	21.15	23.43	29.79	83.45
RTX 5090	14.26	16.44	17.61	15.34	17.25	19.29	25.83	79.17
RTX PRO 6000	19.79	22.72	20.03	20.30	22.02	24.04	30.10	40.93
A100	27.01	28.60	27.06	28.05	30.15	46.83	68.40	109.39
H100	20.75	21.57	20.82	21.69	22.66	25.76	44.05	65.36
H200	19.52	20.33	19.57	20.41	21.43	24.28	38.83	55.49
B200	21.08	22.61	22.93	21.56	24.42	26.81	38.00	63.57
RTX PRO 6000 Workstation × 4	52.96	56.86	51.41	56.51	60.40	60.10	59.86	60.28

Bold: lowest displayed median in each column, including ties.

1,024 history samples (nh)

Fixed platform order across all history lengths and precisions

Platform	np=8	np=16	np=32	np=64	np=96	np=128	np=192	np=256
RTX 5090 (Windows)	22.17	21.76	20.25	21.50	25.24	29.83	111.04	176.65
RTX 5090	16.37	14.93	14.46	17.07	20.92	25.45	107.25	175.07
RTX PRO 6000	22.82	20.18	20.03	22.32	26.05	32.13	118.07	198.83
A100	28.24	30.33	27.66	32.57	45.26	64.51	119.59	190.54
H100	21.62	22.53	21.52	22.72	38.05	42.11	75.46	112.13
H200	20.42	20.96	20.90	21.70	32.12	37.58	59.93	83.29
B200	23.03	22.90	23.85	22.97	32.83	38.11	64.49	98.66
RTX PRO 6000 Workstation × 4	56.54	56.64	57.03	59.98	60.45	60.09	59.64	65.97

Bold: lowest displayed median in each column, including ties.

2,048 history samples (nh)

Fixed platform order across all history lengths and precisions

Platform	np=8	np=16	np=32	np=64	np=96	np=128	np=192	np=256
RTX 5090 (Windows)	20.83	21.53	21.55	24.47	31.62	80.89	200.33	337.73
RTX 5090	14.77	17.00	16.38	20.33	27.56	77.15	197.37	336.88
RTX PRO 6000	19.71	20.71	22.21	26.38	34.15	46.06	231.43	376.38
A100	29.43	31.65	30.83	43.90	68.67	107.57	230.35	346.79
H100	22.20	23.68	23.23	26.92	48.26	67.60	128.92	206.07
H200	20.93	22.73	22.56	25.09	42.30	56.91	96.58	145.18
B200	23.67	24.91	23.09	30.98	43.30	64.63	97.97	162.32
RTX PRO 6000 Workstation × 4	56.60	57.09	59.50	60.13	60.72	59.46	66.85	82.99

Bold: lowest displayed median in each column, including ties.

Platform labels link to their full profiles. Row order is fixed; CV and host-resident timings appear in the platform results.

GPU performance comparison

FP64 · Device-resident median step latency (μs) · Lower is faster

Lower latencyHigher latency

Shared log color scale: 10.00-1,000 μs

512 history samples (nh)

Fixed platform order across all history lengths and precisions

Platform	np=8	np=16	np=32	np=64	np=96	np=128	np=192	np=256
RTX 5090 (Windows)	20.58	20.57	22.45	22.81	25.70	30.43	109.93	179.40
RTX 5090	13.51	14.47	17.49	18.89	21.35	26.82	105.76	175.96
RTX PRO 6000	18.65	19.61	22.70	24.58	27.07	33.04	110.43	203.42
A100	25.85	27.71	25.72	28.43	41.32	59.65	111.15	176.39
H100	21.38	20.45	21.53	21.26	26.54	40.63	70.27	106.12
H200	18.62	19.53	19.94	19.96	23.47	37.45	55.30	79.46
B200	21.77	22.20	21.95	21.93	24.09	24.60	48.74	64.06
RTX PRO 6000 Workstation × 4	51.24	54.40	58.78	59.25	59.23	68.87	64.70	73.48

Bold: lowest displayed median in each column, including ties.

1,024 history samples (nh)

Fixed platform order across all history lengths and precisions

Platform	np=8	np=16	np=32	np=64	np=96	np=128	np=192	np=256
RTX 5090 (Windows)	20.34	22.00	22.07	25.09	34.38	85.16	201.21	336.97
RTX 5090	14.04	16.26	17.28	22.69	29.66	79.78	197.92	338.95
RTX PRO 6000	19.14	21.66	23.44	29.49	39.21	54.61	222.11	379.43
A100	26.79	30.39	27.80	38.71	61.56	94.69	200.72	331.99
H100	19.85	21.96	21.53	25.37	44.69	62.39	119.73	193.06
H200	18.82	20.50	20.24	24.10	39.80	53.69	86.64	140.38
B200	21.20	22.21	21.94	24.55	27.84	42.18	71.58	102.12
RTX PRO 6000 Workstation × 4	54.02	57.87	59.12	59.40	73.12	81.52	70.72	104.63

Bold: lowest displayed median in each column, including ties.

2,048 history samples (nh)

Fixed platform order across all history lengths and precisions

Platform	np=8	np=16	np=32	np=64	np=96	np=128	np=192	np=256
RTX 5090 (Windows)	20.86	23.09	37.26	33.41	112.75	185.11	382.88	650.52
RTX 5090	14.68	19.16	33.07	31.39	109.80	181.86	380.28	655.80
RTX PRO 6000	20.22	24.68	36.47	40.63	119.68	202.38	422.97	731.22
A100	27.97	30.75	31.16	62.25	113.78	177.19	373.46	638.56
H100	21.05	24.78	23.45	41.86	68.39	106.55	216.74	366.92
H200	20.02	23.78	22.79	39.18	52.53	79.20	155.71	256.25
B200	22.01	24.22	24.00	26.33	45.89	65.47	115.75	176.51
RTX PRO 6000 Workstation × 4	58.39	59.09	58.98	69.96	92.17	110.52	155.39	231.63

Bold: lowest displayed median in each column, including ties.

Platform labels link to their full profiles. Row order is fixed; CV and host-resident timings appear in the platform results.

Windows 11 · AMD Ryzen 7 9800X3D (8 CPU cores)

Hardware

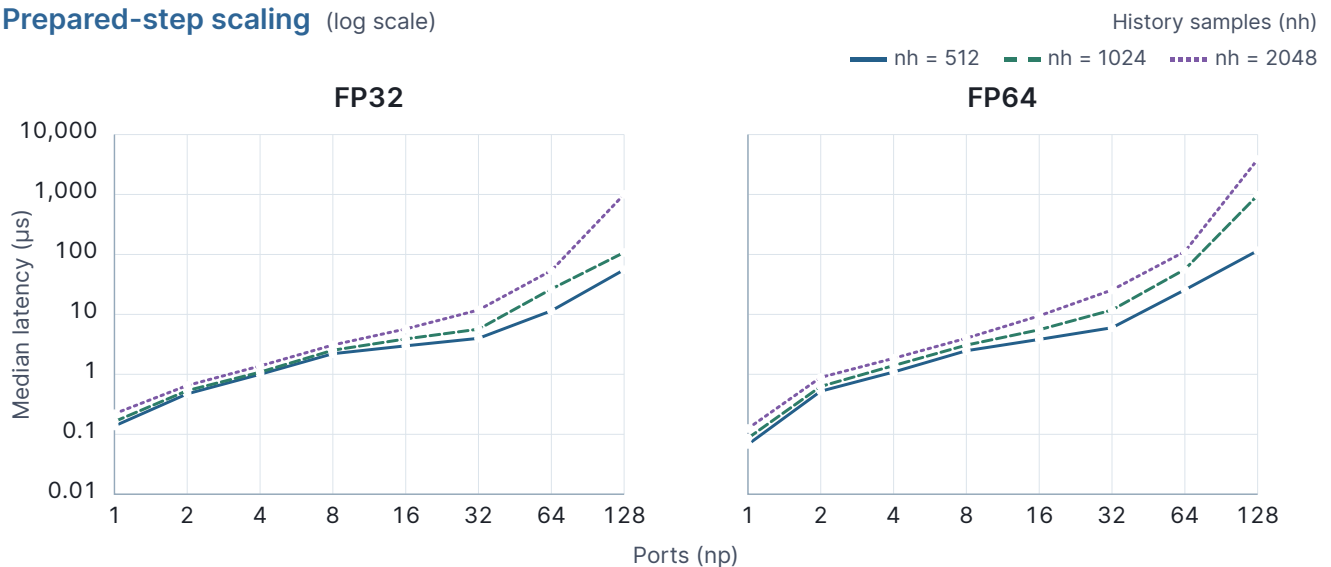
Processor	AMD Ryzen 7 9800X3D
Cores	8 physical cores · SMT enabled
Topology	1 socket · 1 NUMA node
Last-level cache	L3: 1 shared group, 96 MiB
Memory	Total 31.2 GiB
Frequency	Observed snapshot 4700 MHz

Software and measurement

Execution environment	Local workstation
Operating system	Windows 11 · build 10.0.26200
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 5.68 μ s (FP32) and 11.86 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 8; dispatched tasks 1–8 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: 9800X3D · 8 cores. Exact medians and CV for all 48 cases appear on the following page.

Windows 11 · AMD Ryzen 7 9800X3D (8 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.14 μs CV 4.20% · H 0.002 MB	0.16 μs CV 0.49% · H 0.004 MB	0.22 μs CV 0.51% · H 0.008 MB
2	0.47 μs CV 0.61% · H 0.008 MB	0.54 μs CV 0.47% · H 0.02 MB	0.65 μs CV 0.29% · H 0.03 MB
4	1.00 μs CV 0.73% · H 0.03 MB	1.09 μs CV 0.24% · H 0.07 MB	1.38 μs CV 0.53% · H 0.13 MB
8	2.21 μs CV 1.46% · H 0.13 MB	2.52 μs CV 0.86% · H 0.26 MB	3.09 μs CV 0.47% · H 0.52 MB
16	2.99 μs CV 0.94% · H 0.52 MB	3.89 μs CV 1.55% · H 1.05 MB	5.72 μs CV 4.69% · H 2.10 MB
32	3.98 μs CV 1.54% · H 2.10 MB	5.68 μs CV 1.38% · H 4.19 MB	11.85 μs CV 1.92% · H 8.39 MB
64	11.45 μs CV 1.11% · H 8.39 MB	26.47 μs CV 1.65% · H 16.78 MB	53.81 μs CV 0.85% · H 33.55 MB
128	55.58 μs CV 1.11% · H 33.55 MB	107.99 μs CV 2.40% · H 67.11 MB	1,004.55 μs CV 1.53% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.07 μs CV 1.54% · H 0.004 MB	0.09 μs CV 2.11% · H 0.008 MB	0.13 μs CV 0.90% · H 0.02 MB
2	0.52 μs CV 0.81% · H 0.02 MB	0.63 μs CV 0.32% · H 0.03 MB	0.89 μs CV 0.48% · H 0.07 MB
4	1.09 μs CV 0.48% · H 0.07 MB	1.40 μs CV 0.39% · H 0.13 MB	1.85 μs CV 1.02% · H 0.26 MB
8	2.48 μs CV 0.46% · H 0.26 MB	3.08 μs CV 0.66% · H 0.52 MB	4.00 μs CV 1.62% · H 1.05 MB
16	3.82 μs CV 0.85% · H 1.05 MB	5.54 μs CV 0.45% · H 2.10 MB	9.45 μs CV 3.39% · H 4.19 MB
32	5.99 μs CV 3.95% · H 4.19 MB	11.86 μs CV 1.07% · H 8.39 MB	25.51 μs CV 0.82% · H 16.78 MB
64	25.64 μs CV 3.55% · H 16.78 MB	56.35 μs CV 0.98% · H 33.55 MB	111.77 μs CV 0.53% · H 67.11 MB
128	115.72 μs CV 0.69% · H 67.11 MB	968.28 μs CV 0.98% · H 134.22 MB	3,841.08 μs CV 0.54% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c8i.2xlarge · Intel Xeon 6975P-C (4 CPU cores)

Hardware

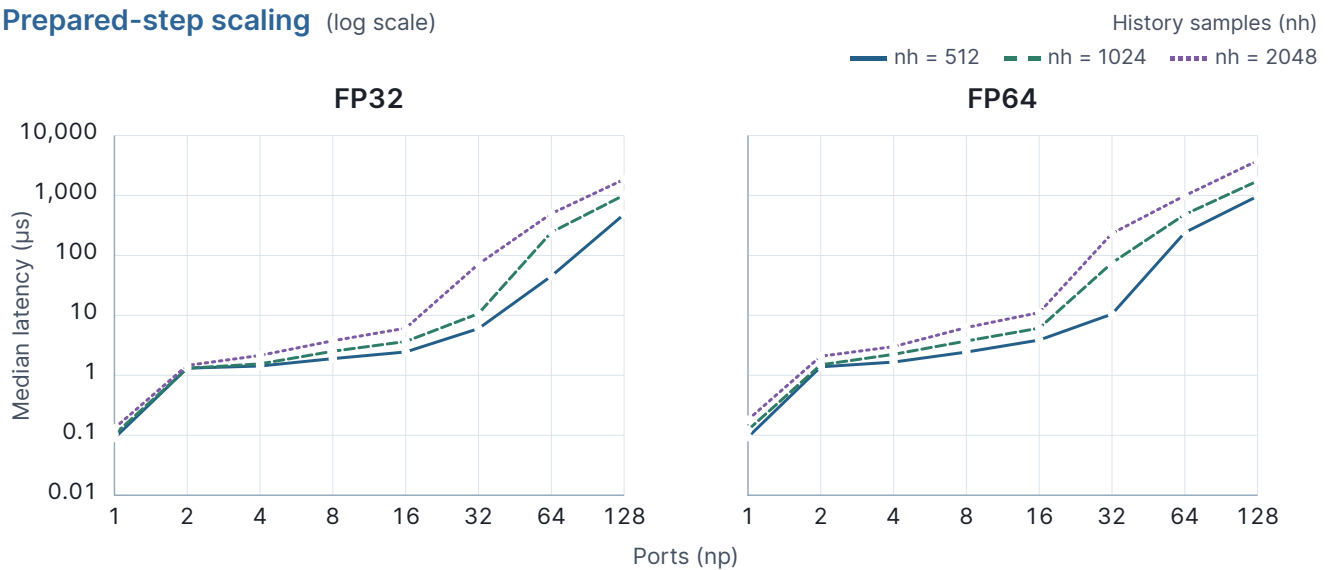
Processor	Intel Xeon 6975P-C
Cores	4 physical cores · SMT enabled
Topology	OS-reported: 1 socket · 1 NUMA node
Last-level cache	L3: 1 shared group, 480 MiB
Memory	Total 15.3 GiB; NUMA node totals 15.3 GiB
Frequency	Observed snapshot 2700 MHz

Software and measurement

Execution environment	AWS c8i.2xlarge (us-east-1 / us-east-1a)
Operating system	Ubuntu 24.04.4 LTS
Kernel	7.0.0-1012-aws
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 10.70 μ s (FP32) and 75.19 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 4; dispatched tasks 1–4 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c8i.2xlarge · 4 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c8i.2xlarge · Intel Xeon 6975P-C (4 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.09 μs CV 7.12% · H 0.002 MB	0.10 μs CV 0.10% · H 0.004 MB	0.13 μs CV 0.11% · H 0.008 MB
2	1.32 μs CV 0.17% · H 0.008 MB	1.31 μs CV 0.10% · H 0.02 MB	1.46 μs CV 0.06% · H 0.03 MB
4	1.43 μs CV 0.20% · H 0.03 MB	1.55 μs CV 0.15% · H 0.07 MB	2.15 μs CV 0.33% · H 0.13 MB
8	1.90 μs CV 0.18% · H 0.13 MB	2.52 μs CV 0.11% · H 0.26 MB	3.78 μs CV 0.03% · H 0.52 MB
16	2.45 μs CV 0.20% · H 0.52 MB	3.67 μs CV 0.07% · H 1.05 MB	6.12 μs CV 0.05% · H 2.10 MB
32	6.03 μs CV 0.12% · H 2.10 MB	10.70 μs CV 0.08% · H 4.19 MB	71.89 μs CV 0.10% · H 8.39 MB
64	45.23 μs CV 0.06% · H 8.39 MB	243.40 μs CV 0.03% · H 16.78 MB	499.04 μs CV 0.07% · H 33.55 MB
128	486.38 μs CV 0.04% · H 33.55 MB	1,014.45 μs CV 0.16% · H 67.11 MB	1,861.05 μs CV 0.97% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.09 μs CV 0.08% · H 0.004 MB	0.12 μs CV 0.10% · H 0.008 MB	0.18 μs CV 0.04% · H 0.02 MB
2	1.38 μs CV 0.20% · H 0.02 MB	1.50 μs CV 0.07% · H 0.03 MB	2.08 μs CV 0.18% · H 0.07 MB
4	1.67 μs CV 0.10% · H 0.07 MB	2.24 μs CV 0.38% · H 0.13 MB	3.02 μs CV 0.22% · H 0.26 MB
8	2.44 μs CV 0.08% · H 0.26 MB	3.75 μs CV 0.17% · H 0.52 MB	6.28 μs CV 0.11% · H 1.05 MB
16	3.88 μs CV 0.13% · H 1.05 MB	6.14 μs CV 0.07% · H 2.10 MB	11.17 μs CV 0.20% · H 4.19 MB
32	10.52 μs CV 0.17% · H 4.19 MB	75.19 μs CV 0.61% · H 8.39 MB	233.28 μs CV 0.20% · H 16.78 MB
64	240.92 μs CV 0.13% · H 16.78 MB	483.77 μs CV 0.15% · H 33.55 MB	997.39 μs CV 0.20% · H 67.11 MB
128	972.84 μs CV 0.18% · H 67.11 MB	1,746.78 μs CV 0.32% · H 134.22 MB	3,811.16 μs CV 0.26% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c8i.4xlarge · Intel Xeon 6975P-C (8 CPU cores)

Hardware

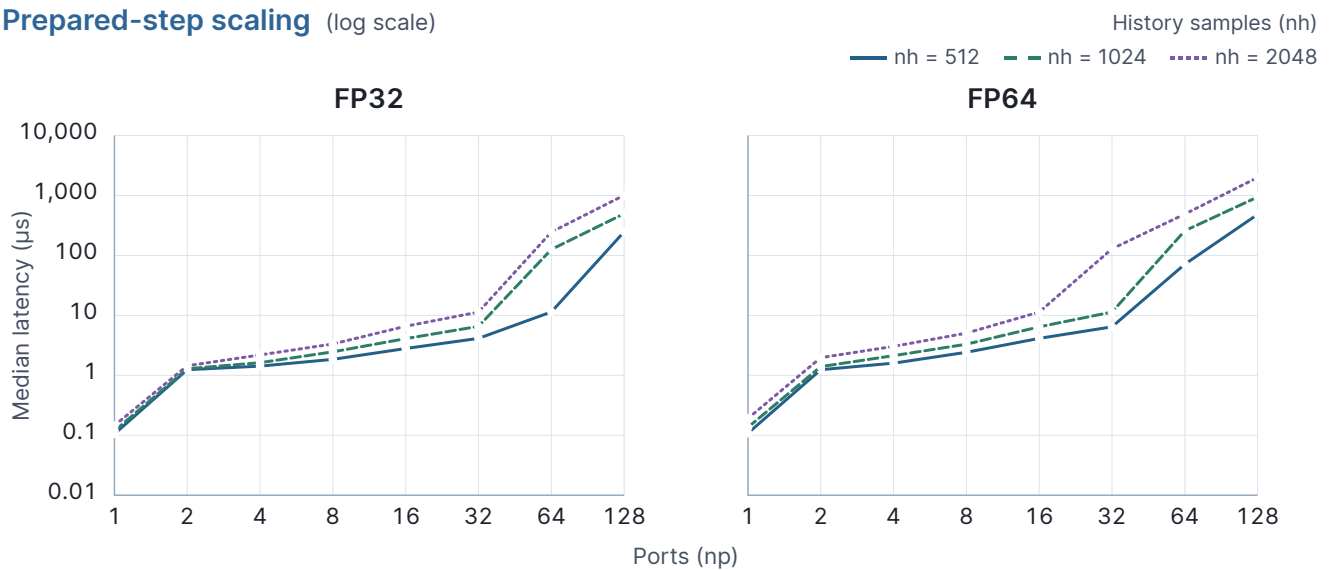
Processor	Intel Xeon 6975P-C
Cores	8 physical cores · one thread per core
Topology	OS-reported: 1 socket · 1 NUMA node
Last-level cache	L3: 1 shared group, 480 MiB
Memory	Total 30.8 GiB; NUMA node totals 30.8 GiB
Frequency	Observed snapshot 3829 MHz

Software and measurement

Execution environment	AWS c8i.4xlarge (us-east-1 / us-east-1a)
Operating system	Ubuntu 24.04.4 LTS
Kernel	7.0.0-1012-aws
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 6.56 μ s (FP32) and 11.35 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 8; dispatched tasks 1–8 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c8i.4xlarge · 8 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c8i.4xlarge · Intel Xeon 6975P-C (8 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.11 μs CV 0.25% · H 0.002 MB	0.12 μs CV 0.22% · H 0.004 MB	0.15 μs CV 0.19% · H 0.008 MB
2	1.24 μs CV 0.06% · H 0.008 MB	1.29 μs CV 0.03% · H 0.02 MB	1.44 μs CV 0.04% · H 0.03 MB
4	1.43 μs CV 0.07% · H 0.03 MB	1.63 μs CV 0.08% · H 0.07 MB	2.19 μs CV 0.10% · H 0.13 MB
8	1.85 μs CV 0.17% · H 0.13 MB	2.47 μs CV 0.09% · H 0.26 MB	3.34 μs CV 0.08% · H 0.52 MB
16	2.80 μs CV 0.18% · H 0.52 MB	4.10 μs CV 0.13% · H 1.05 MB	6.59 μs CV 0.12% · H 2.10 MB
32	4.13 μs CV 0.18% · H 2.10 MB	6.56 μs CV 0.08% · H 4.19 MB	11.37 μs CV 0.29% · H 8.39 MB
64	11.43 μs CV 0.17% · H 8.39 MB	126.33 μs CV 0.04% · H 16.78 MB	246.40 μs CV 0.02% · H 33.55 MB
128	248.12 μs CV 0.11% · H 33.55 MB	498.25 μs CV 0.02% · H 67.11 MB	1,010.43 μs CV 0.09% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.11 μs CV 0.29% · H 0.004 MB	0.14 μs CV 0.22% · H 0.008 MB	0.19 μs CV 0.15% · H 0.02 MB
2	1.24 μs CV 0.10% · H 0.02 MB	1.40 μs CV 0.10% · H 0.03 MB	1.97 μs CV 0.10% · H 0.07 MB
4	1.59 μs CV 0.06% · H 0.07 MB	2.13 μs CV 0.04% · H 0.13 MB	3.05 μs CV 0.10% · H 0.26 MB
8	2.42 μs CV 0.15% · H 0.26 MB	3.31 μs CV 0.08% · H 0.52 MB	5.05 μs CV 0.09% · H 1.05 MB
16	4.14 μs CV 0.10% · H 1.05 MB	6.45 μs CV 0.33% · H 2.10 MB	11.39 μs CV 0.32% · H 4.19 MB
32	6.46 μs CV 0.09% · H 4.19 MB	11.35 μs CV 0.10% · H 8.39 MB	129.22 μs CV 0.12% · H 16.78 MB
64	71.64 μs CV 0.35% · H 16.78 MB	258.23 μs CV 0.06% · H 33.55 MB	490.55 μs CV 0.02% · H 67.11 MB
128	484.47 μs CV 0.04% · H 67.11 MB	945.08 μs CV 0.03% · H 134.22 MB	1,994.11 μs CV 0.06% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c8i.8xlarge · Intel Xeon 6975P-C (16 CPU cores)

Hardware

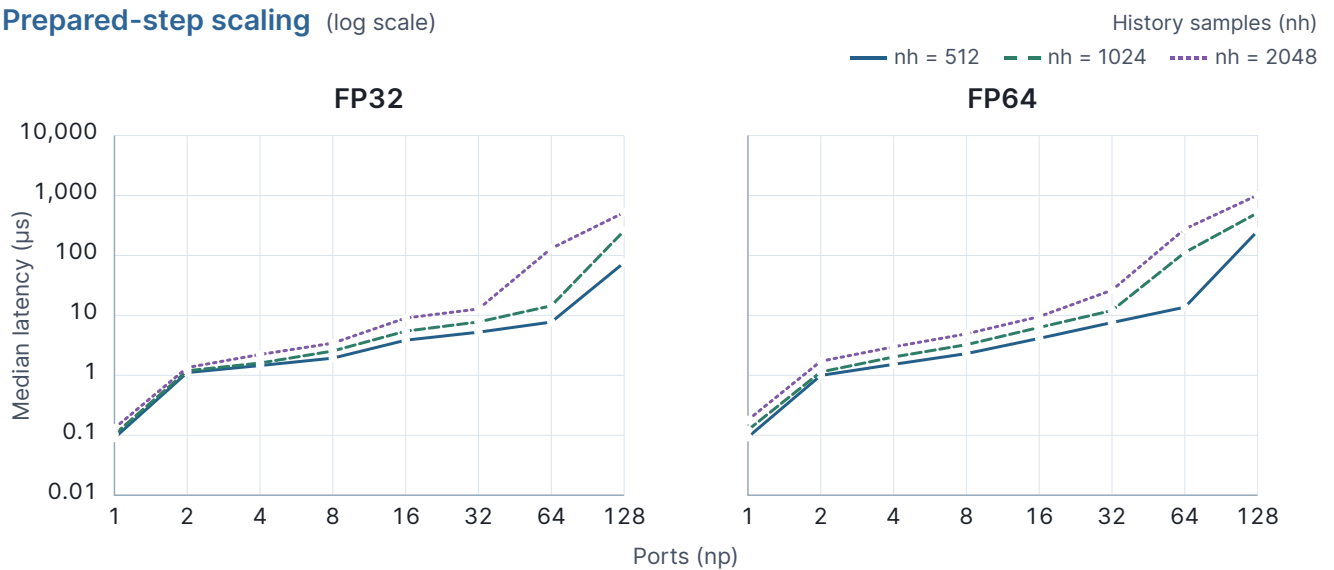
Processor	Intel Xeon 6975P-C
Cores	16 physical cores · one thread per core
Topology	OS-reported: 1 socket · 1 NUMA node
Last-level cache	L3: 1 shared group, 480 MiB
Memory	Total 61.8 GiB; NUMA node totals 61.8 GiB
Frequency	Observed snapshot 3898 MHz

Software and measurement

Execution environment	AWS c8i.8xlarge (us-east-1 / us-east-1a)
Operating system	Ubuntu 24.04.4 LTS
Kernel	7.0.0-1012-aws
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 7.78 μ s (FP32) and 12.06 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 16; dispatched tasks 1–16 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c8i.8xlarge · 16 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c8i.8xlarge · Intel Xeon 6975P-C (16 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.09 μs CV 0.08% · H 0.002 MB	0.10 μs CV 0.12% · H 0.004 MB	0.13 μs CV 0.10% · H 0.008 MB
2	1.11 μs CV 0.13% · H 0.008 MB	1.19 μs CV 0.18% · H 0.02 MB	1.35 μs CV 0.05% · H 0.03 MB
4	1.46 μs CV 0.40% · H 0.03 MB	1.60 μs CV 0.13% · H 0.07 MB	2.23 μs CV 0.19% · H 0.13 MB
8	1.94 μs CV 0.08% · H 0.13 MB	2.54 μs CV 0.16% · H 0.26 MB	3.45 μs CV 0.10% · H 0.52 MB
16	3.86 μs CV 0.10% · H 0.52 MB	5.46 μs CV 0.13% · H 1.05 MB	9.06 μs CV 0.11% · H 2.10 MB
32	5.24 μs CV 0.14% · H 2.10 MB	7.78 μs CV 0.12% · H 4.19 MB	12.78 μs CV 0.08% · H 8.39 MB
64	7.74 μs CV 0.04% · H 8.39 MB	14.50 μs CV 0.04% · H 16.78 MB	131.48 μs CV 0.07% · H 33.55 MB
128	76.00 μs CV 0.03% · H 33.55 MB	258.90 μs CV 0.10% · H 67.11 MB	521.60 μs CV 0.25% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.09 μs CV 0.06% · H 0.004 MB	0.12 μs CV 0.26% · H 0.008 MB	0.18 μs CV 0.10% · H 0.02 MB
2	1.00 μs CV 0.12% · H 0.02 MB	1.14 μs CV 0.44% · H 0.03 MB	1.72 μs CV 0.10% · H 0.07 MB
4	1.52 μs CV 0.36% · H 0.07 MB	2.02 μs CV 0.18% · H 0.13 MB	2.99 μs CV 0.07% · H 0.26 MB
8	2.30 μs CV 0.10% · H 0.26 MB	3.25 μs CV 0.14% · H 0.52 MB	4.92 μs CV 0.18% · H 1.05 MB
16	4.14 μs CV 0.20% · H 1.05 MB	6.30 μs CV 0.39% · H 2.10 MB	9.55 μs CV 0.06% · H 4.19 MB
32	7.59 μs CV 0.10% · H 4.19 MB	12.06 μs CV 0.05% · H 8.39 MB	26.36 μs CV 0.09% · H 16.78 MB
64	13.70 μs CV 0.11% · H 16.78 MB	112.48 μs CV 0.20% · H 33.55 MB	277.30 μs CV 0.03% · H 67.11 MB
128	256.52 μs CV 0.16% · H 67.11 MB	513.31 μs CV 0.03% · H 134.22 MB	1,014.00 μs CV 0.42% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c8i.metal-48xl · Intel Xeon 6975P-C (96 CPU cores)

Hardware

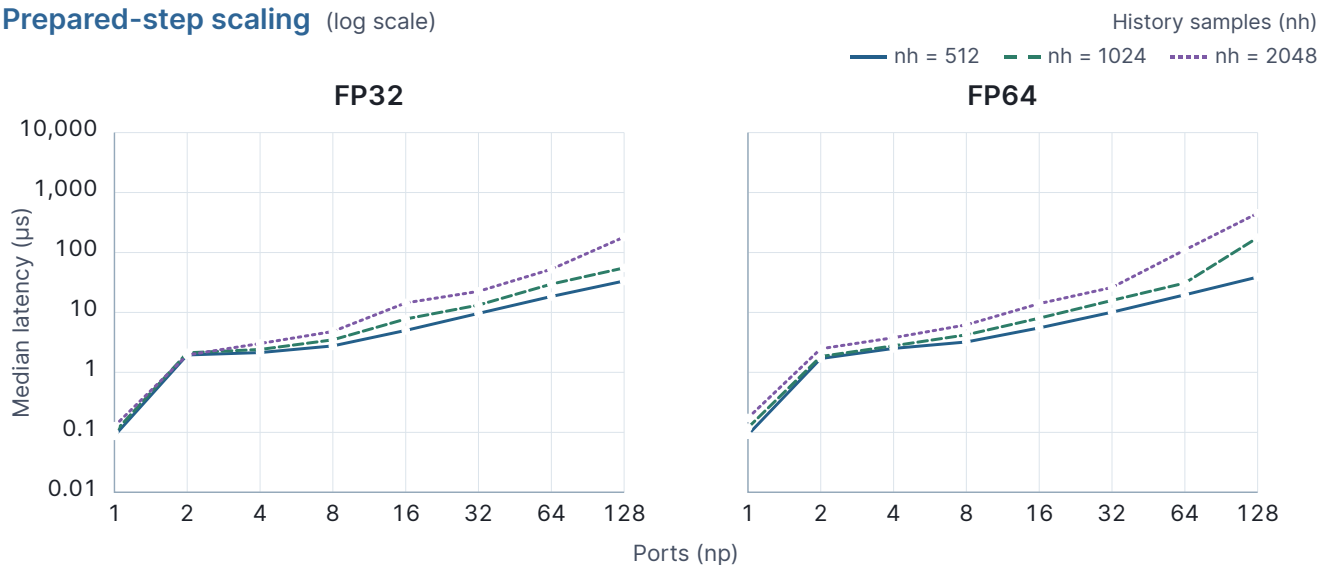
Processor	Intel Xeon 6975P-C
Cores	96 physical cores · SMT enabled
Topology	OS-reported: 1 socket · 3 NUMA nodes
Last-level cache	L3: 1 shared group, 480 MiB
Memory	Total 377.6 GiB; NUMA node totals 125.7 GiB, 125.9 GiB, 126.0 GiB
Frequency	Observed snapshot 3663 MHz; policy range 800-3900 MHz; governor performance

Software and measurement

Execution environment	AWS c8i.metal-48xl (us-east-1 / us-east-1a)
Operating system	Ubuntu 24.04.4 LTS
Kernel	7.0.0-1012-aws
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 13.29 μ s (FP32) and 15.90 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 96; dispatched tasks 1-96 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c8i.metal-48xl · 96 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c8i.metal-48xl · Intel Xeon 6975P-C (96 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.09 μs CV 0.98% · H 0.002 MB	0.10 μs CV 0.08% · H 0.004 MB	0.13 μs CV 0.08% · H 0.008 MB
2	1.96 μs CV 0.04% · H 0.008 MB	2.12 μs CV 1.32% · H 0.02 MB	1.95 μs CV 0.04% · H 0.03 MB
4	2.14 μs CV 1.66% · H 0.03 MB	2.41 μs CV 2.54% · H 0.07 MB	3.02 μs CV 0.61% · H 0.13 MB
8	2.77 μs CV 0.07% · H 0.13 MB	3.49 μs CV 0.06% · H 0.26 MB	4.80 μs CV 0.05% · H 0.52 MB
16	5.01 μs CV 0.13% · H 0.52 MB	7.76 μs CV 0.07% · H 1.05 MB	14.41 μs CV 0.03% · H 2.10 MB
32	9.66 μs CV 0.09% · H 2.10 MB	13.29 μs CV 0.07% · H 4.19 MB	22.52 μs CV 0.03% · H 8.39 MB
64	18.42 μs CV 0.06% · H 8.39 MB	29.66 μs CV 0.05% · H 16.78 MB	52.52 μs CV 0.02% · H 33.55 MB
128	33.45 μs CV 0.06% · H 33.55 MB	55.28 μs CV 0.04% · H 67.11 MB	181.20 μs CV 0.18% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.09 μs CV 0.21% · H 0.004 MB	0.12 μs CV 0.23% · H 0.008 MB	0.17 μs CV 0.17% · H 0.02 MB
2	1.70 μs CV 0.08% · H 0.02 MB	1.84 μs CV 0.09% · H 0.03 MB	2.47 μs CV 0.07% · H 0.07 MB
4	2.51 μs CV 0.05% · H 0.07 MB	2.78 μs CV 0.09% · H 0.13 MB	3.79 μs CV 0.09% · H 0.26 MB
8	3.22 μs CV 0.14% · H 0.26 MB	4.23 μs CV 0.17% · H 0.52 MB	6.21 μs CV 0.07% · H 1.05 MB
16	5.51 μs CV 0.04% · H 1.05 MB	8.02 μs CV 0.09% · H 2.10 MB	14.01 μs CV 0.04% · H 4.19 MB
32	10.09 μs CV 0.05% · H 4.19 MB	15.90 μs CV 0.05% · H 8.39 MB	26.12 μs CV 0.03% · H 16.78 MB
64	19.71 μs CV 0.06% · H 16.78 MB	30.89 μs CV 0.03% · H 33.55 MB	110.55 μs CV 0.02% · H 67.11 MB
128	38.83 μs CV 0.04% · H 67.11 MB	176.58 μs CV 0.02% · H 134.22 MB	452.07 μs CV 0.04% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c8a.2xlarge · AMD EPYC 9R45 (4 CPU cores)

Hardware

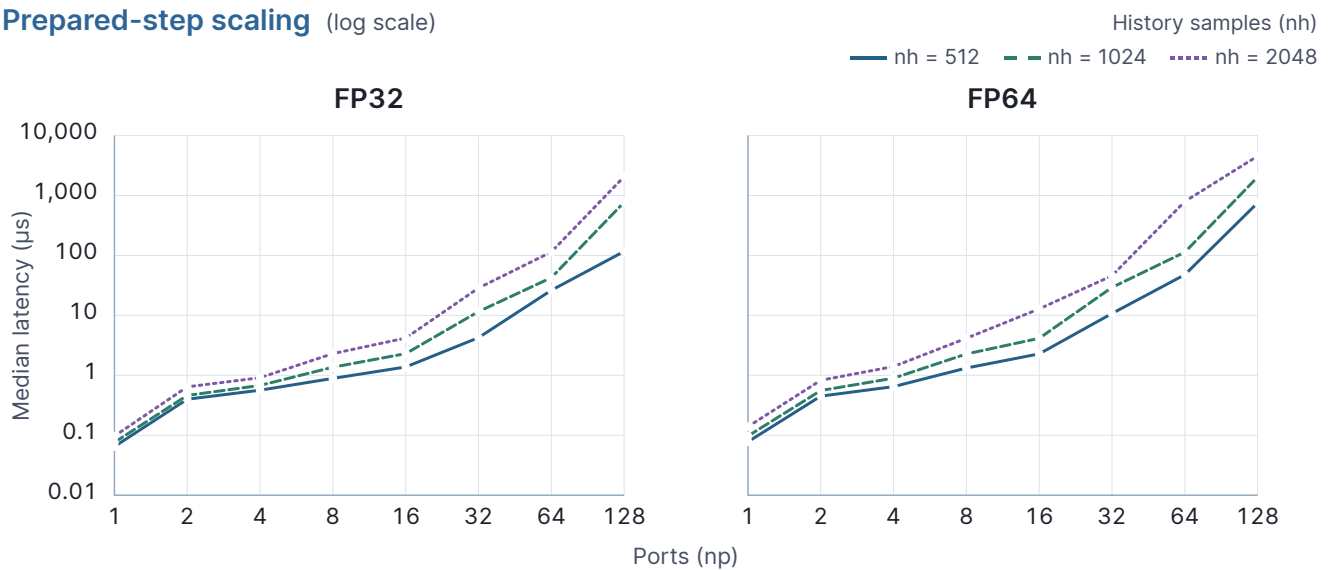
Processor	AMD EPYC 9R45
Cores	4 physical cores · one thread per core
Topology	OS-reported: 1 socket · 1 NUMA node
Last-level cache	L3: 1 shared group, 16 MiB
Memory	Total 15.3 GiB; NUMA node totals 15.3 GiB
Frequency	Observed snapshot 2600 MHz

Software and measurement

Execution environment	AWS c8a.2xlarge (us-east-1 / us-east-1a)
Operating system	Ubuntu 24.04.4 LTS
Kernel	7.0.0-1012-aws
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 11.59 μ s (FP32) and 29.26 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 4; dispatched tasks 1–4 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c8a.2xlarge · 4 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c8a.2xlarge · AMD EPYC 9R45 (4 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.06 μs CV 0.14% · H 0.002 MB	0.08 μs CV 0.13% · H 0.004 MB	0.10 μs CV 0.13% · H 0.008 MB
2	0.40 μs CV 1.58% · H 0.008 MB	0.46 μs CV 0.09% · H 0.02 MB	0.64 μs CV 0.09% · H 0.03 MB
4	0.57 μs CV 0.24% · H 0.03 MB	0.68 μs CV 0.18% · H 0.07 MB	0.92 μs CV 0.43% · H 0.13 MB
8	0.89 μs CV 0.34% · H 0.13 MB	1.37 μs CV 0.18% · H 0.26 MB	2.31 μs CV 0.08% · H 0.52 MB
16	1.38 μs CV 0.08% · H 0.52 MB	2.31 μs CV 0.17% · H 1.05 MB	4.19 μs CV 0.13% · H 2.10 MB
32	4.25 μs CV 0.16% · H 2.10 MB	11.59 μs CV 0.41% · H 4.19 MB	28.64 μs CV 0.07% · H 8.39 MB
64	26.23 μs CV 0.17% · H 8.39 MB	42.71 μs CV 0.04% · H 16.78 MB	114.60 μs CV 0.68% · H 33.55 MB
128	116.69 μs CV 2.41% · H 33.55 MB	771.68 μs CV 0.38% · H 67.11 MB	2,059.03 μs CV 0.09% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.08 μs CV 0.90% · H 0.004 MB	0.10 μs CV 0.11% · H 0.008 MB	0.14 μs CV 0.10% · H 0.02 MB
2	0.45 μs CV 1.25% · H 0.02 MB	0.56 μs CV 0.38% · H 0.03 MB	0.83 μs CV 0.11% · H 0.07 MB
4	0.65 μs CV 0.23% · H 0.07 MB	0.90 μs CV 0.09% · H 0.13 MB	1.40 μs CV 0.05% · H 0.26 MB
8	1.33 μs CV 0.08% · H 0.26 MB	2.27 μs CV 0.10% · H 0.52 MB	4.15 μs CV 0.12% · H 1.05 MB
16	2.29 μs CV 0.09% · H 1.05 MB	4.15 μs CV 0.60% · H 2.10 MB	12.83 μs CV 0.50% · H 4.19 MB
32	10.75 μs CV 0.52% · H 4.19 MB	29.26 μs CV 0.15% · H 8.39 MB	45.95 μs CV 0.16% · H 16.78 MB
64	47.76 μs CV 0.16% · H 16.78 MB	113.36 μs CV 0.50% · H 33.55 MB	795.67 μs CV 0.21% · H 67.11 MB
128	752.98 μs CV 0.37% · H 67.11 MB	2,030.65 μs CV 0.23% · H 134.22 MB	4,562.18 μs CV 0.38% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c8a.2xlarge · AMD EPYC 9R45 (8 CPU cores)

Hardware

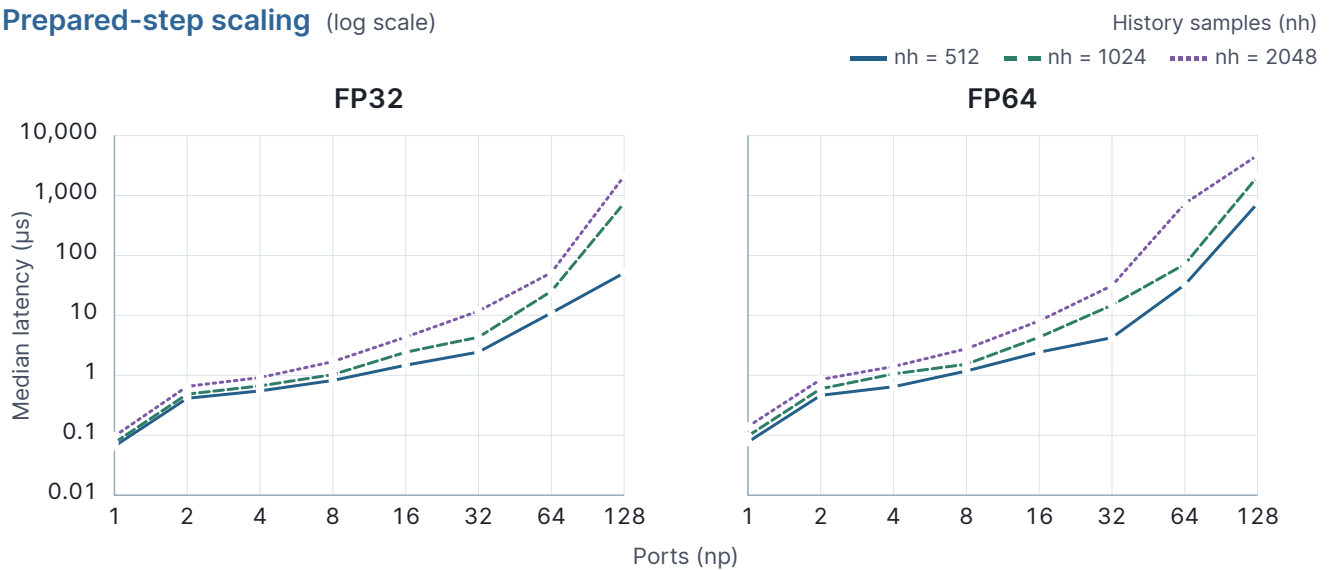
Processor	AMD EPYC 9R45
Cores	8 physical cores · one thread per core
Topology	OS-reported: 1 socket · 1 NUMA node
Last-level cache	L3: 1 shared group, 32 MiB
Memory	Total 15.3 GiB; NUMA node totals 15.3 GiB
Frequency	Observed snapshot 4461 MHz

Software and measurement

Execution environment	AWS c8a.2xlarge (us-east-1 / us-east-1a)
Operating system	Ubuntu 24.04.4 LTS
Kernel	7.0.0-1012-aws
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 4.32 μ s (FP32) and 14.83 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 8; dispatched tasks 1–8 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c8a.2xlarge · 8 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c8a.2xlarge · AMD EPYC 9R45 (8 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.07 μs CV 0.18% · H 0.002 MB	0.08 μs CV 0.10% · H 0.004 MB	0.10 μs CV 0.11% · H 0.008 MB
2	0.41 μs CV 0.19% · H 0.008 MB	0.49 μs CV 0.14% · H 0.02 MB	0.65 μs CV 0.13% · H 0.03 MB
4	0.55 μs CV 0.15% · H 0.03 MB	0.67 μs CV 0.16% · H 0.07 MB	0.92 μs CV 0.20% · H 0.13 MB
8	0.82 μs CV 0.10% · H 0.13 MB	1.03 μs CV 0.09% · H 0.26 MB	1.70 μs CV 0.05% · H 0.52 MB
16	1.48 μs CV 0.12% · H 0.52 MB	2.43 μs CV 0.07% · H 1.05 MB	4.36 μs CV 0.03% · H 2.10 MB
32	2.45 μs CV 0.02% · H 2.10 MB	4.32 μs CV 0.03% · H 4.19 MB	11.78 μs CV 0.04% · H 8.39 MB
64	11.01 μs CV 0.03% · H 8.39 MB	25.27 μs CV 0.05% · H 16.78 MB	51.71 μs CV 0.36% · H 33.55 MB
128	50.93 μs CV 0.04% · H 33.55 MB	748.91 μs CV 0.18% · H 67.11 MB	2,101.55 μs CV 0.51% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.08 μs CV 0.22% · H 0.004 MB	0.10 μs CV 0.11% · H 0.008 MB	0.14 μs CV 0.10% · H 0.02 MB
2	0.47 μs CV 1.23% · H 0.02 MB	0.60 μs CV 0.21% · H 0.03 MB	0.85 μs CV 0.43% · H 0.07 MB
4	0.64 μs CV 0.08% · H 0.07 MB	1.06 μs CV 0.11% · H 0.13 MB	1.41 μs CV 0.17% · H 0.26 MB
8	1.18 μs CV 0.14% · H 0.26 MB	1.53 μs CV 0.16% · H 0.52 MB	2.78 μs CV 0.12% · H 1.05 MB
16	2.43 μs CV 0.07% · H 1.05 MB	4.33 μs CV 0.68% · H 2.10 MB	8.11 μs CV 0.09% · H 4.19 MB
32	4.26 μs CV 0.37% · H 4.19 MB	14.83 μs CV 0.27% · H 8.39 MB	31.39 μs CV 0.24% · H 16.78 MB
64	32.57 μs CV 0.85% · H 16.78 MB	71.47 μs CV 0.47% · H 33.55 MB	737.98 μs CV 0.13% · H 67.11 MB
128	743.91 μs CV 0.45% · H 67.11 MB	2,077.68 μs CV 0.26% · H 134.22 MB	4,701.03 μs CV 0.18% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c8a.4xlarge · AMD EPYC 9R45 (16 CPU cores)

Hardware

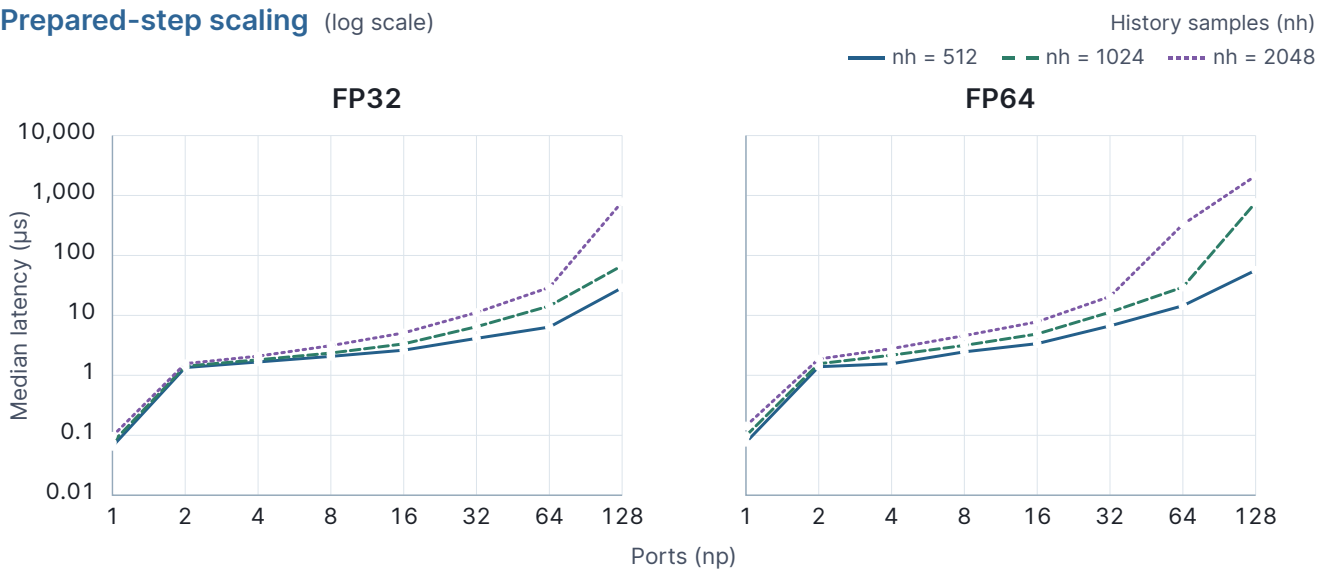
Processor	AMD EPYC 9R45
Cores	16 physical cores · one thread per core
Topology	OS-reported: 1 socket · 1 NUMA node
Last-level cache	L3: 2 × 32 MiB shared groups
Memory	Total 30.7 GiB; NUMA node totals 30.7 GiB
Frequency	Observed snapshot 2600 MHz

Software and measurement

Execution environment	AWS c8a.4xlarge (us-east-1 / us-east-1a)
Operating system	Ubuntu 24.04.4 LTS
Kernel	7.0.0-1012-aws
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 6.48 μ s (FP32) and 11.32 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 16; dispatched tasks 1–16 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c8a.4xlarge · 16 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c8a.4xlarge · AMD EPYC 9R45 (16 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.07 μs CV 0.21% · H 0.002 MB	0.08 μs CV 0.10% · H 0.004 MB	0.10 μs CV 0.07% · H 0.008 MB
2	1.35 μs CV 0.29% · H 0.008 MB	1.42 μs CV 0.39% · H 0.02 MB	1.57 μs CV 0.26% · H 0.03 MB
4	1.67 μs CV 0.26% · H 0.03 MB	1.81 μs CV 0.49% · H 0.07 MB	2.10 μs CV 0.32% · H 0.13 MB
8	2.07 μs CV 0.16% · H 0.13 MB	2.35 μs CV 0.11% · H 0.26 MB	3.12 μs CV 0.15% · H 0.52 MB
16	2.63 μs CV 0.08% · H 0.52 MB	3.35 μs CV 0.11% · H 1.05 MB	5.10 μs CV 0.08% · H 2.10 MB
32	4.14 μs CV 0.19% · H 2.10 MB	6.48 μs CV 0.08% · H 4.19 MB	11.17 μs CV 0.71% · H 8.39 MB
64	6.39 μs CV 0.11% · H 8.39 MB	14.37 μs CV 0.54% · H 16.78 MB	29.35 μs CV 0.44% · H 33.55 MB
128	28.91 μs CV 0.45% · H 33.55 MB	67.56 μs CV 0.37% · H 67.11 MB	785.77 μs CV 0.10% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.08 μs CV 0.09% · H 0.004 MB	0.10 μs CV 0.05% · H 0.008 MB	0.14 μs CV 0.04% · H 0.02 MB
2	1.39 μs CV 0.12% · H 0.02 MB	1.56 μs CV 0.07% · H 0.03 MB	1.87 μs CV 0.18% · H 0.07 MB
4	1.56 μs CV 0.26% · H 0.07 MB	2.17 μs CV 0.12% · H 0.13 MB	2.79 μs CV 0.21% · H 0.26 MB
8	2.47 μs CV 0.25% · H 0.26 MB	3.16 μs CV 0.09% · H 0.52 MB	4.60 μs CV 0.11% · H 1.05 MB
16	3.38 μs CV 0.07% · H 1.05 MB	4.90 μs CV 0.25% · H 2.10 MB	7.80 μs CV 0.05% · H 4.19 MB
32	6.68 μs CV 0.45% · H 4.19 MB	11.32 μs CV 0.05% · H 8.39 MB	20.74 μs CV 0.14% · H 16.78 MB
64	14.49 μs CV 0.10% · H 16.78 MB	29.52 μs CV 0.14% · H 33.55 MB	329.98 μs CV 0.10% · H 67.11 MB
128	56.70 μs CV 0.31% · H 67.11 MB	763.78 μs CV 0.03% · H 134.22 MB	2,129.87 μs CV 0.02% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c8a.metal-24xl · AMD EPYC 9R45 (96 CPU cores)

Hardware

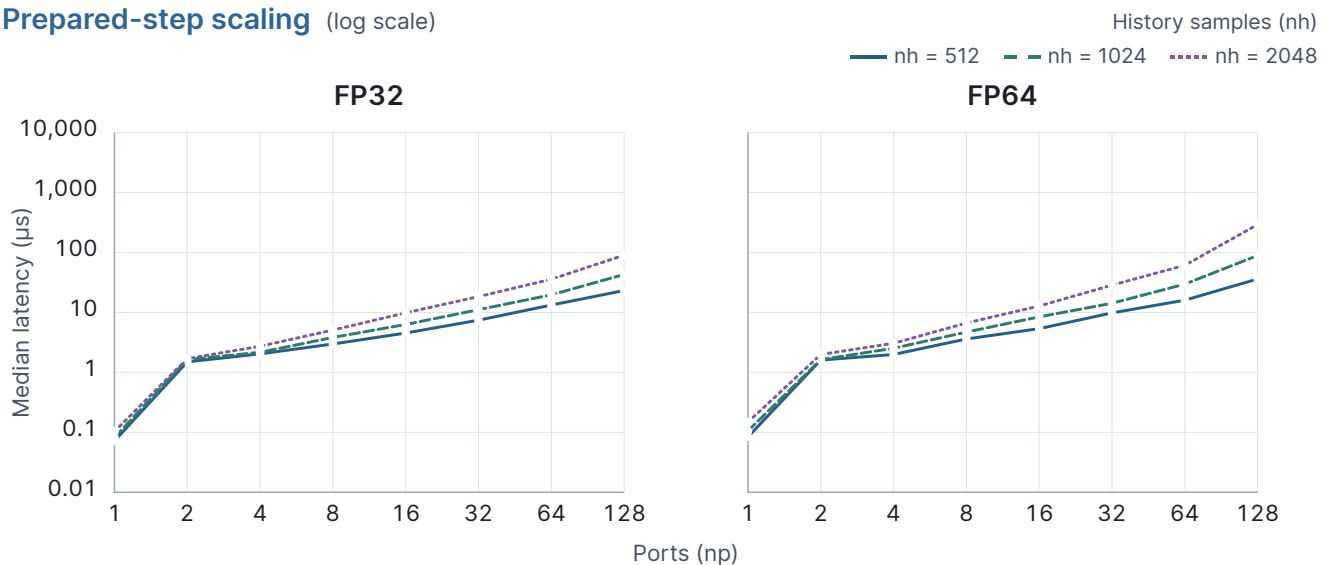
Processor	AMD EPYC 9R45
Cores	96 physical cores · one thread per core
Topology	OS-reported: 1 socket · 1 NUMA node
Last-level cache	L3: 12 × 32 MiB shared groups
Memory	Total 187.5 GiB; NUMA node totals 187.5 GiB
Frequency	Observed snapshot 1999 MHz; policy range 1999-4510 MHz; governor performance

Software and measurement

Execution environment	AWS c8a.metal-24xl (us-east-1 / us-east-1a)
Operating system	Ubuntu 24.04.4 LTS
Kernel	7.0.0-1012-aws
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 11.11 μ s (FP32) and 14.25 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 96; dispatched tasks 1-96 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c8a.metal-24xl · 96 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c8a.metal-24xl · AMD EPYC 9R45 (96 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.07 μs CV 0.40% · H 0.002 MB	0.08 μs CV 0.20% · H 0.004 MB	0.10 μs CV 0.03% · H 0.008 MB
2	1.50 μs CV 0.23% · H 0.008 MB	1.64 μs CV 0.78% · H 0.02 MB	1.69 μs CV 0.04% · H 0.03 MB
4	2.03 μs CV 0.11% · H 0.03 MB	2.16 μs CV 0.25% · H 0.07 MB	2.72 μs CV 0.25% · H 0.13 MB
8	2.98 μs CV 0.68% · H 0.13 MB	3.83 μs CV 0.23% · H 0.26 MB	5.08 μs CV 0.49% · H 0.52 MB
16	4.53 μs CV 0.15% · H 0.52 MB	6.30 μs CV 0.15% · H 1.05 MB	9.85 μs CV 0.20% · H 2.10 MB
32	7.43 μs CV 0.14% · H 2.10 MB	11.11 μs CV 0.11% · H 4.19 MB	18.39 μs CV 0.16% · H 8.39 MB
64	13.26 μs CV 0.19% · H 8.39 MB	19.70 μs CV 0.18% · H 16.78 MB	35.77 μs CV 0.11% · H 33.55 MB
128	23.24 μs CV 0.27% · H 33.55 MB	43.01 μs CV 0.06% · H 67.11 MB	90.24 μs CV 0.07% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.08 μs CV 0.61% · H 0.004 MB	0.11 μs CV 0.08% · H 0.008 MB	0.15 μs CV 0.19% · H 0.02 MB
2	1.61 μs CV 0.51% · H 0.02 MB	1.64 μs CV 0.30% · H 0.03 MB	1.99 μs CV 0.36% · H 0.07 MB
4	1.99 μs CV 0.25% · H 0.07 MB	2.52 μs CV 0.32% · H 0.13 MB	3.06 μs CV 0.41% · H 0.26 MB
8	3.61 μs CV 0.21% · H 0.26 MB	4.69 μs CV 0.10% · H 0.52 MB	6.66 μs CV 0.25% · H 1.05 MB
16	5.38 μs CV 0.13% · H 1.05 MB	8.43 μs CV 0.20% · H 2.10 MB	12.82 μs CV 0.26% · H 4.19 MB
32	9.80 μs CV 0.18% · H 4.19 MB	14.25 μs CV 0.11% · H 8.39 MB	28.57 μs CV 0.12% · H 16.78 MB
64	16.04 μs CV 0.12% · H 16.78 MB	29.75 μs CV 0.24% · H 33.55 MB	61.13 μs CV 0.05% · H 67.11 MB
128	36.11 μs CV 0.07% · H 67.11 MB	89.00 μs CV 0.26% · H 134.22 MB	292.61 μs CV 0.20% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c7i.4xlarge · Intel Xeon Platinum 8488C (8 CPU cores)

Hardware

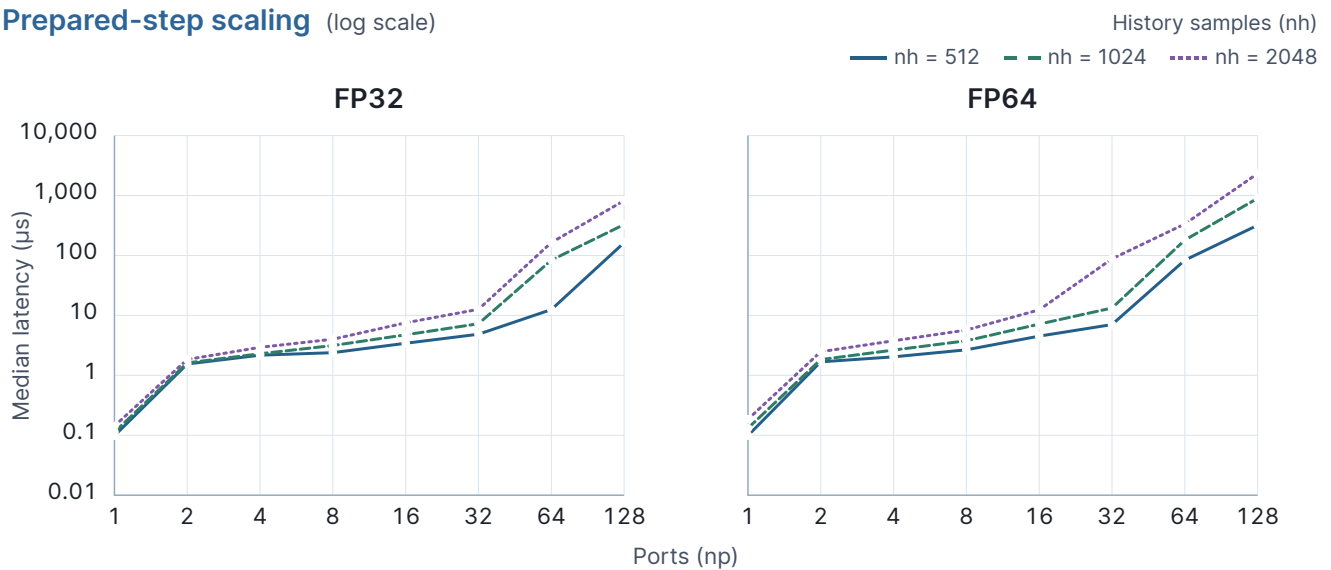
Processor	Intel Xeon Platinum 8488C
Cores	8 physical cores · one thread per core
Topology	OS-reported: 1 socket · 1 NUMA node
Last-level cache	L3: 1 shared group, 105 MiB
Memory	Total 30.8 GiB; NUMA node totals 30.8 GiB
Frequency	Observed snapshot 2400 MHz

Software and measurement

Execution environment	AWS c7i.4xlarge (us-east-1 / us-east-1a)
Operating system	Ubuntu 24.04.4 LTS
Kernel	7.0.0-1012-aws
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 7.26 μ s (FP32) and 13.26 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 8; dispatched tasks 1–8 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c7i.4xlarge · 8 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c7i.4xlarge · Intel Xeon Platinum 8488C (8 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.10 μs CV 0.46% · H 0.002 MB	0.11 μs CV 0.84% · H 0.004 MB	0.14 μs CV 0.79% · H 0.008 MB
2	1.55 μs CV 0.09% · H 0.008 MB	1.61 μs CV 0.11% · H 0.02 MB	1.84 μs CV 0.19% · H 0.03 MB
4	2.17 μs CV 0.51% · H 0.03 MB	2.30 μs CV 0.17% · H 0.07 MB	2.94 μs CV 0.63% · H 0.13 MB
8	2.39 μs CV 1.10% · H 0.13 MB	3.14 μs CV 0.38% · H 0.26 MB	4.00 μs CV 0.55% · H 0.52 MB
16	3.43 μs CV 0.55% · H 0.52 MB	4.75 μs CV 0.79% · H 1.05 MB	7.52 μs CV 0.64% · H 2.10 MB
32	4.85 μs CV 0.55% · H 2.10 MB	7.26 μs CV 0.70% · H 4.19 MB	12.63 μs CV 0.82% · H 8.39 MB
64	12.54 μs CV 3.15% · H 8.39 MB	83.48 μs CV 0.14% · H 16.78 MB	164.50 μs CV 0.68% · H 33.55 MB
128	163.55 μs CV 0.08% · H 33.55 MB	327.36 μs CV 0.06% · H 67.11 MB	824.50 μs CV 0.40% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.10 μs CV 0.24% · H 0.004 MB	0.13 μs CV 0.49% · H 0.008 MB	0.19 μs CV 0.32% · H 0.02 MB
2	1.68 μs CV 0.17% · H 0.02 MB	1.83 μs CV 0.74% · H 0.03 MB	2.47 μs CV 0.18% · H 0.07 MB
4	2.04 μs CV 0.23% · H 0.07 MB	2.64 μs CV 0.63% · H 0.13 MB	3.78 μs CV 0.23% · H 0.26 MB
8	2.67 μs CV 0.96% · H 0.26 MB	3.77 μs CV 0.70% · H 0.52 MB	5.70 μs CV 0.58% · H 1.05 MB
16	4.52 μs CV 3.56% · H 1.05 MB	7.16 μs CV 1.25% · H 2.10 MB	12.40 μs CV 0.43% · H 4.19 MB
32	7.04 μs CV 1.00% · H 4.19 MB	13.26 μs CV 0.28% · H 8.39 MB	87.97 μs CV 0.06% · H 16.78 MB
64	83.38 μs CV 0.05% · H 16.78 MB	181.36 μs CV 0.14% · H 33.55 MB	336.09 μs CV 0.05% · H 67.11 MB
128	318.61 μs CV 0.06% · H 67.11 MB	896.72 μs CV 1.30% · H 134.22 MB	2,328.40 μs CV 1.75% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c7a.2xlarge · AMD EPYC 9R14 (8 CPU cores)

Hardware

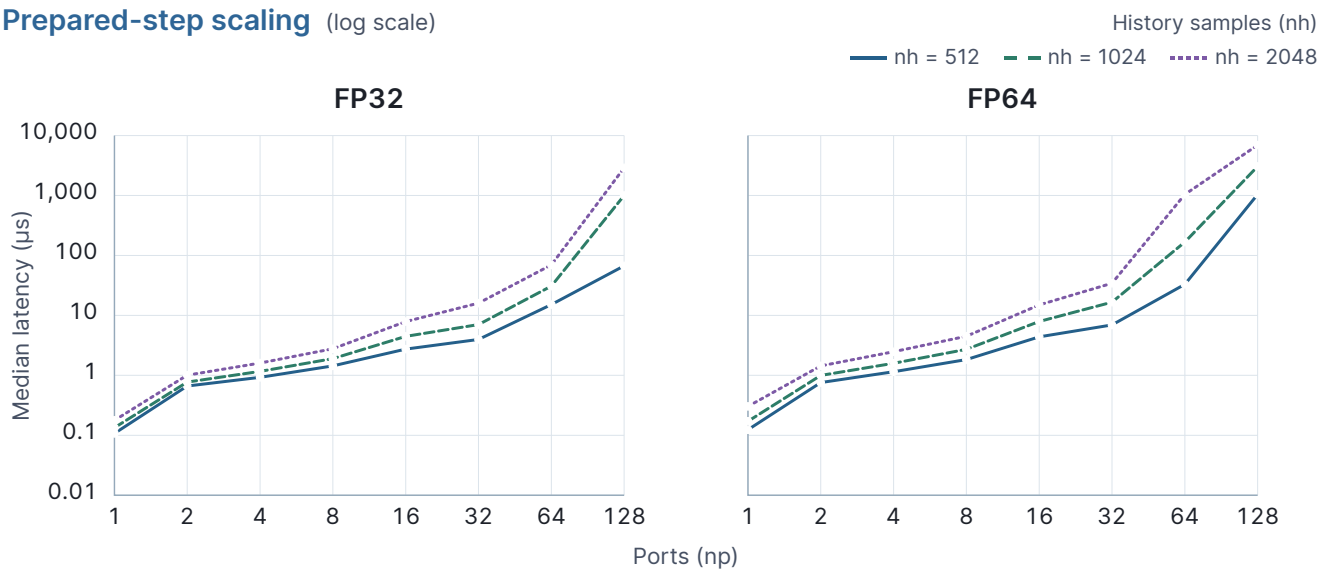
Processor	AMD EPYC 9R14
Cores	8 physical cores · one thread per core
Topology	OS-reported: 1 socket · 1 NUMA node
Last-level cache	L3: 1 shared group, 32 MiB
Memory	Total 15.2 GiB; NUMA node totals 15.2 GiB
Frequency	Observed snapshot 2600 MHz

Software and measurement

Execution environment	AWS c7a.2xlarge (us-east-1 / us-east-1a)
Operating system	Ubuntu 24.04.4 LTS
Kernel	7.0.0-1012-aws
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 7.04 μ s (FP32) and 16.63 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 8; dispatched tasks 1–8 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c7a.2xlarge · 8 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c7a.2xlarge · AMD EPYC 9R14 (8 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.11 μs CV 0.87% · H 0.002 MB	0.13 μs CV 0.21% · H 0.004 MB	0.18 μs CV 0.04% · H 0.008 MB
2	0.67 μs CV 0.05% · H 0.008 MB	0.77 μs CV 0.05% · H 0.02 MB	1.01 μs CV 0.02% · H 0.03 MB
4	0.93 μs CV 0.08% · H 0.03 MB	1.16 μs CV 0.04% · H 0.07 MB	1.60 μs CV 0.25% · H 0.13 MB
8	1.44 μs CV 0.89% · H 0.13 MB	1.90 μs CV 0.12% · H 0.26 MB	2.78 μs CV 0.27% · H 0.52 MB
16	2.73 μs CV 0.17% · H 0.52 MB	4.48 μs CV 0.63% · H 1.05 MB	7.87 μs CV 0.13% · H 2.10 MB
32	3.97 μs CV 0.26% · H 2.10 MB	7.04 μs CV 0.13% · H 4.19 MB	16.01 μs CV 0.95% · H 8.39 MB
64	15.15 μs CV 0.65% · H 8.39 MB	30.80 μs CV 1.00% · H 16.78 MB	69.62 μs CV 0.46% · H 33.55 MB
128	67.20 μs CV 0.31% · H 33.55 MB	1,002.47 μs CV 0.10% · H 67.11 MB	2,901.14 μs CV 0.11% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.12 μs CV 0.06% · H 0.004 MB	0.17 μs CV 0.06% · H 0.008 MB	0.30 μs CV 0.10% · H 0.02 MB
2	0.76 μs CV 0.04% · H 0.02 MB	1.00 μs CV 0.03% · H 0.03 MB	1.44 μs CV 0.12% · H 0.07 MB
4	1.15 μs CV 0.09% · H 0.07 MB	1.59 μs CV 0.07% · H 0.13 MB	2.48 μs CV 0.35% · H 0.26 MB
8	1.83 μs CV 1.04% · H 0.26 MB	2.71 μs CV 0.72% · H 0.52 MB	4.48 μs CV 1.05% · H 1.05 MB
16	4.38 μs CV 0.13% · H 1.05 MB	7.90 μs CV 0.66% · H 2.10 MB	14.97 μs CV 0.67% · H 4.19 MB
32	6.95 μs CV 0.32% · H 4.19 MB	16.63 μs CV 1.08% · H 8.39 MB	34.41 μs CV 0.33% · H 16.78 MB
64	33.05 μs CV 0.28% · H 16.78 MB	167.25 μs CV 4.10% · H 33.55 MB	1,041.99 μs CV 0.25% · H 67.11 MB
128	1,042.15 μs CV 0.57% · H 67.11 MB	3,062.74 μs CV 0.08% · H 134.22 MB	6,928.73 μs CV 0.09% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c9g.2xlarge · aarch64 (8 CPU cores)

Hardware

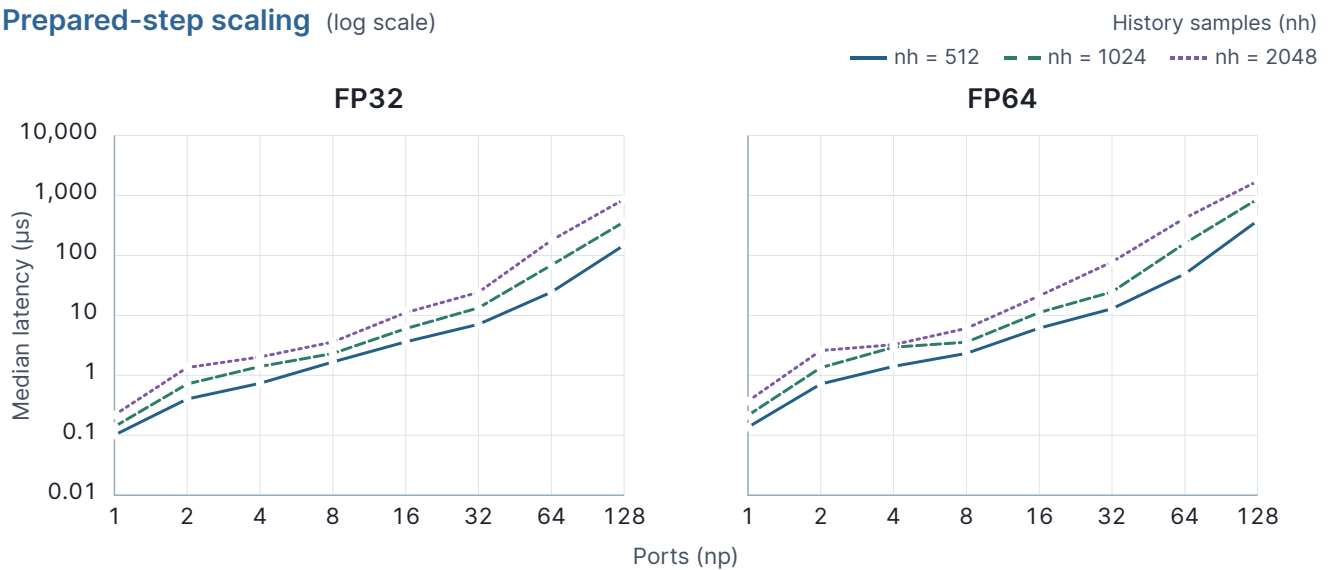
Processor	aarch64
Cores	8 physical cores · one thread per core
Topology	OS-reported: 1 socket · 1 NUMA node
Last-level cache	L3: 1 shared group, 48 MiB
Memory	Total 15.3 GiB; NUMA node totals 15.3 GiB
Frequency	Not captured

Software and measurement

Execution environment	AWS c9g.2xlarge (us-east-1 / us-east-1a)
Operating system	Ubuntu 24.04.4 LTS
Kernel	7.0.0-1012-aws
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 13.33 μ s (FP32) and 24.67 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 8; dispatched tasks 1–8 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c9g.2xlarge · 8 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c9g.2xlarge · aarch64 (8 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.10 μs CV 0.04% · H 0.002 MB	0.14 μs CV 0.02% · H 0.004 MB	0.22 μs CV 0.02% · H 0.008 MB
2	0.40 μs CV 0.02% · H 0.008 MB	0.72 μs CV 0.05% · H 0.02 MB	1.35 μs CV 0.02% · H 0.03 MB
4	0.74 μs CV 0.01% · H 0.03 MB	1.40 μs CV 0.10% · H 0.07 MB	2.02 μs CV 0.06% · H 0.13 MB
8	1.67 μs CV 0.08% · H 0.13 MB	2.32 μs CV 0.03% · H 0.26 MB	3.61 μs CV 0.04% · H 0.52 MB
16	3.63 μs CV 0.05% · H 0.52 MB	6.03 μs CV 0.03% · H 1.05 MB	11.12 μs CV 0.46% · H 2.10 MB
32	7.08 μs CV 0.05% · H 2.10 MB	13.33 μs CV 0.04% · H 4.19 MB	24.28 μs CV 0.05% · H 8.39 MB
64	24.42 μs CV 0.07% · H 8.39 MB	69.00 μs CV 0.07% · H 16.78 MB	178.43 μs CV 0.06% · H 33.55 MB
128	148.80 μs CV 0.05% · H 33.55 MB	362.01 μs CV 0.04% · H 67.11 MB	863.20 μs CV 0.24% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.14 μs CV 0.04% · H 0.004 MB	0.21 μs CV 0.04% · H 0.008 MB	0.37 μs CV 0.02% · H 0.02 MB
2	0.72 μs CV 0.05% · H 0.02 MB	1.34 μs CV 0.03% · H 0.03 MB	2.60 μs CV 0.05% · H 0.07 MB
4	1.41 μs CV 0.15% · H 0.07 MB	2.95 μs CV 0.01% · H 0.13 MB	3.25 μs CV 0.04% · H 0.26 MB
8	2.32 μs CV 0.04% · H 0.26 MB	3.57 μs CV 0.01% · H 0.52 MB	6.07 μs CV 0.02% · H 1.05 MB
16	6.12 μs CV 0.05% · H 1.05 MB	11.14 μs CV 0.02% · H 2.10 MB	21.17 μs CV 0.20% · H 4.19 MB
32	12.91 μs CV 0.16% · H 4.19 MB	24.67 μs CV 0.04% · H 8.39 MB	77.92 μs CV 0.12% · H 16.78 MB
64	49.13 μs CV 0.14% · H 16.78 MB	159.48 μs CV 0.24% · H 33.55 MB	420.00 μs CV 0.05% · H 67.11 MB
128	379.52 μs CV 0.04% · H 67.11 MB	874.95 μs CV 0.21% · H 134.22 MB	1,743.84 μs CV 0.16% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c9g.4xlarge · aarch64 (16 CPU cores)

Hardware

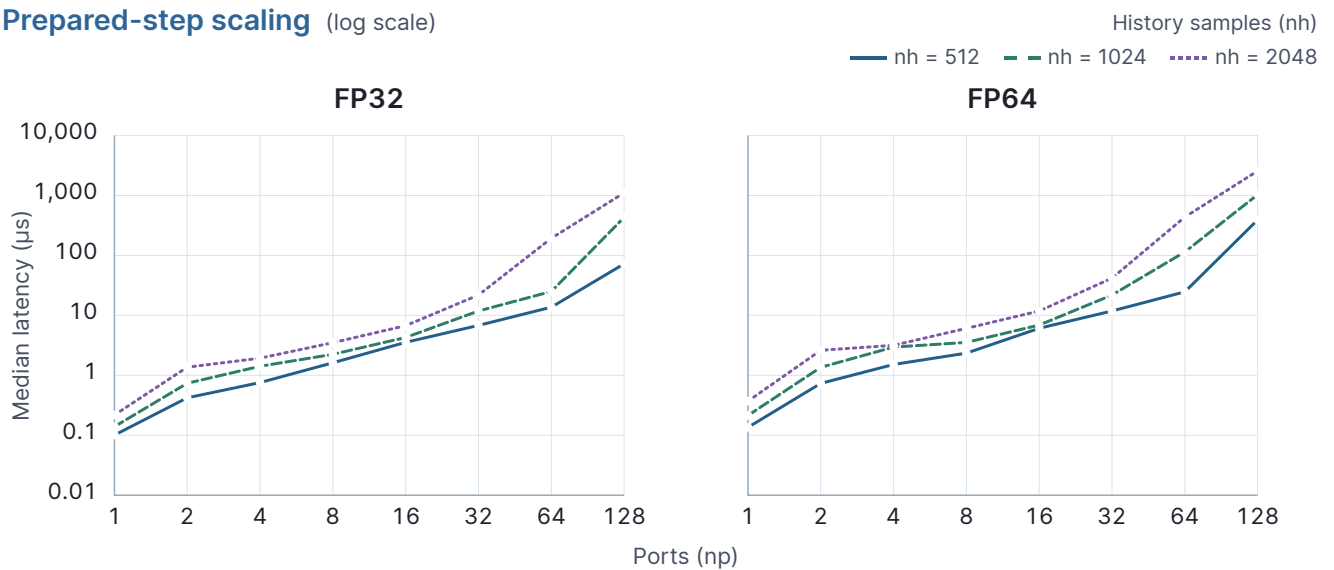
Processor	aarch64
Cores	16 physical cores · one thread per core
Topology	OS-reported: 1 socket · 1 NUMA node
Last-level cache	L3: 1 shared group, 48 MiB
Memory	Total 30.7 GiB; NUMA node totals 30.7 GiB
Frequency	Not captured

Software and measurement

Execution environment	AWS c9g.4xlarge (us-east-1 / us-east-1a)
Operating system	Ubuntu 24.04.4 LTS
Kernel	7.0.0-1012-aws
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 11.82 μ s (FP32) and 21.87 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 16; dispatched tasks 1–16 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c9g.4xlarge · 16 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c9g.4xlarge · aarch64 (16 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.10 μs CV 0.05% · H 0.002 MB	0.14 μs CV 0.03% · H 0.004 MB	0.22 μs CV 0.02% · H 0.008 MB
2	0.42 μs CV 0.03% · H 0.008 MB	0.74 μs CV 0.03% · H 0.02 MB	1.37 μs CV 0.04% · H 0.03 MB
4	0.76 μs CV 0.01% · H 0.03 MB	1.42 μs CV 0.43% · H 0.07 MB	1.94 μs CV 0.02% · H 0.13 MB
8	1.62 μs CV 0.04% · H 0.13 MB	2.24 μs CV 0.14% · H 0.26 MB	3.51 μs CV 0.12% · H 0.52 MB
16	3.56 μs CV 0.02% · H 0.52 MB	4.29 μs CV 0.04% · H 1.05 MB	6.76 μs CV 0.05% · H 2.10 MB
32	6.80 μs CV 0.02% · H 2.10 MB	11.82 μs CV 0.06% · H 4.19 MB	21.82 μs CV 0.03% · H 8.39 MB
64	13.72 μs CV 0.09% · H 8.39 MB	25.05 μs CV 0.10% · H 16.78 MB	194.16 μs CV 0.26% · H 33.55 MB
128	72.35 μs CV 0.34% · H 33.55 MB	432.31 μs CV 0.08% · H 67.11 MB	1,108.26 μs CV 0.46% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.14 μs CV 0.06% · H 0.004 MB	0.21 μs CV 0.02% · H 0.008 MB	0.37 μs CV 0.03% · H 0.02 MB
2	0.73 μs CV 0.04% · H 0.02 MB	1.36 μs CV 0.02% · H 0.03 MB	2.62 μs CV 0.02% · H 0.07 MB
4	1.52 μs CV 0.12% · H 0.07 MB	2.96 μs CV 0.02% · H 0.13 MB	3.17 μs CV 0.02% · H 0.26 MB
8	2.35 μs CV 0.03% · H 0.26 MB	3.54 μs CV 0.03% · H 0.52 MB	6.08 μs CV 0.02% · H 1.05 MB
16	6.09 μs CV 0.04% · H 1.05 MB	6.91 μs CV 0.16% · H 2.10 MB	11.84 μs CV 0.03% · H 4.19 MB
32	11.87 μs CV 0.05% · H 4.19 MB	21.87 μs CV 0.04% · H 8.39 MB	41.94 μs CV 0.25% · H 16.78 MB
64	24.76 μs CV 0.16% · H 16.78 MB	114.45 μs CV 0.15% · H 33.55 MB	437.67 μs CV 0.26% · H 67.11 MB
128	392.06 μs CV 0.11% · H 67.11 MB	1,032.21 μs CV 0.52% · H 134.22 MB	2,567.15 μs CV 0.18% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c8g.2xlarge · aarch64 (8 CPU cores)

Hardware

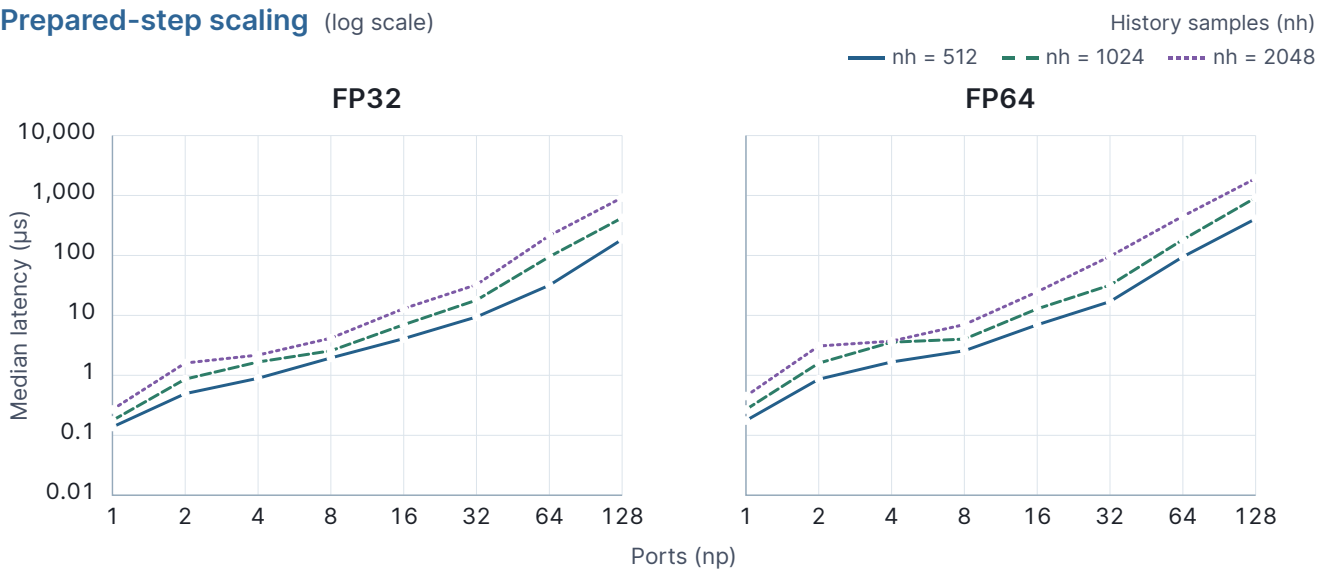
Processor	aarch64
Cores	8 physical cores · one thread per core
Topology	OS-reported: 1 socket · 1 NUMA node
Last-level cache	L3: 1 shared group, 36 MiB
Memory	Total 15.3 GiB; NUMA node totals 15.3 GiB
Frequency	Not captured

Software and measurement

Execution environment	AWS c8g.2xlarge (us-east-1 / us-east-1a)
Operating system	Ubuntu 24.04.4 LTS
Kernel	7.0.0-1012-aws
Math library	OpenBLAS 0.3.34
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 17.99 μ s (FP32) and 31.91 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 8; dispatched tasks 1–8 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c8g.2xlarge · 8 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c8g.2xlarge · aarch64 (8 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.14 μs CV 0.02% · H 0.002 MB	0.18 μs CV 0.05% · H 0.004 MB	0.27 μs CV 0.02% · H 0.008 MB
2	0.49 μs CV 0.02% · H 0.008 MB	0.87 μs CV 0.09% · H 0.02 MB	1.60 μs CV 0.03% · H 0.03 MB
4	0.89 μs CV 0.03% · H 0.03 MB	1.67 μs CV 0.07% · H 0.07 MB	2.18 μs CV 0.02% · H 0.13 MB
8	1.95 μs CV 0.09% · H 0.13 MB	2.57 μs CV 0.30% · H 0.26 MB	4.16 μs CV 0.05% · H 0.52 MB
16	4.09 μs CV 0.07% · H 0.52 MB	6.99 μs CV 0.05% · H 1.05 MB	12.98 μs CV 0.04% · H 2.10 MB
32	9.41 μs CV 0.24% · H 2.10 MB	17.99 μs CV 0.06% · H 4.19 MB	32.49 μs CV 0.03% · H 8.39 MB
64	31.85 μs CV 0.01% · H 8.39 MB	95.21 μs CV 0.18% · H 16.78 MB	214.71 μs CV 0.18% · H 33.55 MB
128	185.15 μs CV 0.18% · H 33.55 MB	423.94 μs CV 0.08% · H 67.11 MB	920.74 μs CV 0.06% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.18 μs CV 0.03% · H 0.004 MB	0.27 μs CV 0.04% · H 0.008 MB	0.45 μs CV 0.02% · H 0.02 MB
2	0.86 μs CV 0.08% · H 0.02 MB	1.60 μs CV 0.05% · H 0.03 MB	3.08 μs CV 0.04% · H 0.07 MB
4	1.67 μs CV 0.31% · H 0.07 MB	3.57 μs CV 0.04% · H 0.13 MB	3.72 μs CV 0.02% · H 0.26 MB
8	2.58 μs CV 0.06% · H 0.26 MB	4.02 μs CV 0.02% · H 0.52 MB	7.04 μs CV 0.19% · H 1.05 MB
16	6.97 μs CV 0.02% · H 1.05 MB	12.86 μs CV 0.04% · H 2.10 MB	24.74 μs CV 0.03% · H 4.19 MB
32	16.98 μs CV 0.07% · H 4.19 MB	31.91 μs CV 0.09% · H 8.39 MB	96.82 μs CV 0.72% · H 16.78 MB
64	94.97 μs CV 0.55% · H 16.78 MB	183.80 μs CV 0.37% · H 33.55 MB	450.82 μs CV 0.09% · H 67.11 MB
128	409.49 μs CV 0.14% · H 67.11 MB	919.09 μs CV 0.08% · H 134.22 MB	1,930.25 μs CV 0.09% · H 268.44 MB

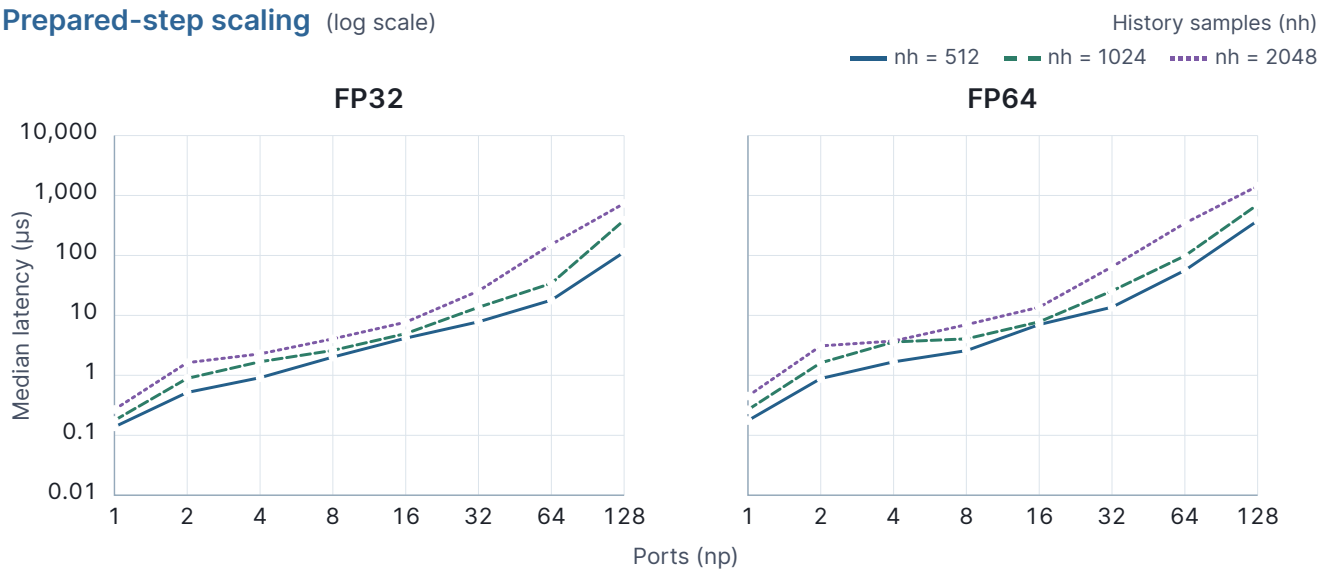
Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS c8g.4xlarge · aarch64 (16 CPU cores)

Hardware		Software and measurement	
Processor	aarch64	Execution environment	AWS c8g.4xlarge (us-east-1 / us-east-1a)
Cores	16 physical cores · one thread per core	Operating system	Ubuntu 24.04.4 LTS
Topology	OS-reported: 1 socket · 1 NUMA node	Kernel	7.0.0-1012-aws
Last-level cache	L3: 1 shared group, 36 MiB	Math library	OpenBLAS 0.3.34
Memory	Total 30.7 GiB; NUMA node totals 30.7 GiB	Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Frequency	Not captured	Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 13.67 μ s (FP32) and 25.55 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

BLAS configured 16; dispatched tasks 1–16 by shape; workers pinned to physical cores. Active concurrency not measured.

Chart label: c8g.4xlarge · 16 cores. Exact medians and CV for all 48 cases appear on the following page.

AWS c8g.4xlarge · aarch64 (16 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.14 μs CV 0.01% · H 0.002 MB	0.18 μs CV 0.04% · H 0.004 MB	0.27 μs CV 0.17% · H 0.008 MB
2	0.52 μs CV 0.04% · H 0.008 MB	0.89 μs CV 0.07% · H 0.02 MB	1.63 μs CV 0.08% · H 0.03 MB
4	0.92 μs CV 0.02% · H 0.03 MB	1.68 μs CV 0.06% · H 0.07 MB	2.28 μs CV 0.01% · H 0.13 MB
8	2.02 μs CV 0.04% · H 0.13 MB	2.60 μs CV 0.03% · H 0.26 MB	4.06 μs CV 0.03% · H 0.52 MB
16	4.17 μs CV 0.03% · H 0.52 MB	4.97 μs CV 0.04% · H 1.05 MB	7.70 μs CV 0.05% · H 2.10 MB
32	7.80 μs CV 0.08% · H 2.10 MB	13.67 μs CV 0.05% · H 4.19 MB	25.50 μs CV 0.05% · H 8.39 MB
64	17.96 μs CV 0.06% · H 8.39 MB	34.26 μs CV 0.04% · H 16.78 MB	153.07 μs CV 0.16% · H 33.55 MB
128	113.74 μs CV 0.12% · H 33.55 MB	386.72 μs CV 0.16% · H 67.11 MB	737.54 μs CV 0.10% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.18 μs CV 0.03% · H 0.004 MB	0.27 μs CV 0.05% · H 0.008 MB	0.45 μs CV 0.03% · H 0.02 MB
2	0.88 μs CV 0.09% · H 0.02 MB	1.62 μs CV 0.08% · H 0.03 MB	3.10 μs CV 0.03% · H 0.07 MB
4	1.68 μs CV 0.14% · H 0.07 MB	3.60 μs CV 0.12% · H 0.13 MB	3.74 μs CV 0.03% · H 0.26 MB
8	2.58 μs CV 0.04% · H 0.26 MB	4.06 μs CV 0.02% · H 0.52 MB	7.01 μs CV 0.06% · H 1.05 MB
16	7.03 μs CV 0.06% · H 1.05 MB	7.81 μs CV 0.26% · H 2.10 MB	13.68 μs CV 0.09% · H 4.19 MB
32	13.73 μs CV 0.03% · H 4.19 MB	25.55 μs CV 0.04% · H 8.39 MB	64.54 μs CV 0.13% · H 16.78 MB
64	56.28 μs CV 1.83% · H 16.78 MB	98.31 μs CV 1.07% · H 33.55 MB	344.18 μs CV 0.11% · H 67.11 MB
128	379.26 μs CV 0.14% · H 67.11 MB	705.11 μs CV 0.06% · H 134.22 MB	1,438.69 μs CV 0.10% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

macOS 27.0 · Apple M5 (10 CPU cores)

Hardware

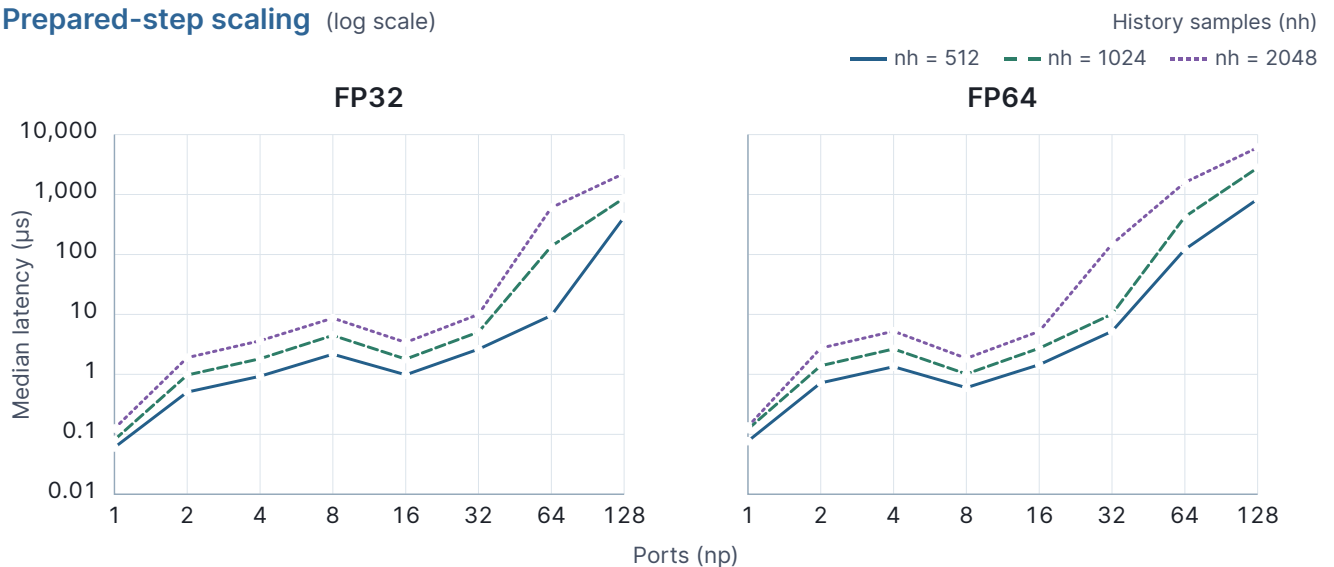
Processor	Apple M5
Cores	10 physical cores (4 Super + 6 Efficiency cores) · one thread per core
Topology	1 socket · single UMA domain · CPU placement managed by macOS
Last-level cache	L2: 1 shared group, 6 MiB
Memory	Total 16.0 GiB
Frequency	Per-core frequency is not exposed by macOS

Software and measurement

Execution environment	Local workstation
Operating system	macOS 27.0
Kernel	27.0.0
Math library	Apple Accelerate framework (GEMV and threading)
Test coverage	48 cases · FP32 and FP64 · np 1-128 · nh 512-2048
Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps

Reference workload. At np=32, nh=1024, TDSE Runtime completes the prepared step in 5.07 μ s (FP32) and 10.11 μ s (FP64).

Prepared-step scaling (log scale)



Timed threading and placement

Requested cap 10; Accelerate auto threading; worker count not observable; macOS scheduler placement, no pinning.

Chart label: Apple M5 · 10 cores. Exact medians and CV for all 48 cases appear on the following page.

macOS 27.0 · Apple M5 (10 CPU cores)

Complete result matrices

FP32 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.06 μs CV 0.25% · H 0.002 MB	0.08 μs CV 0.46% · H 0.004 MB	0.13 μs CV 0.42% · H 0.008 MB
2	0.51 μs CV 0.20% · H 0.008 MB	0.97 μs CV 0.19% · H 0.02 MB	1.90 μs CV 0.07% · H 0.03 MB
4	0.93 μs CV 1.20% · H 0.03 MB	1.82 μs CV 0.59% · H 0.07 MB	3.64 μs CV 0.08% · H 0.13 MB
8	2.17 μs CV 0.14% · H 0.13 MB	4.46 μs CV 0.11% · H 0.26 MB	8.68 μs CV 0.12% · H 0.52 MB
16	0.98 μs CV 0.30% · H 0.52 MB	1.80 μs CV 0.18% · H 1.05 MB	3.42 μs CV 0.24% · H 2.10 MB
32	2.63 μs CV 1.52% · H 2.10 MB	5.07 μs CV 0.38% · H 4.19 MB	9.98 μs CV 0.41% · H 8.39 MB
64	9.61 μs CV 0.86% · H 8.39 MB	137.79 μs CV 0.75% · H 16.78 MB	606.07 μs CV 10.06% · H 33.55 MB
128	420.50 μs CV 0.24% · H 33.55 MB	873.13 μs CV 0.66% · H 67.11 MB	2,251.36 μs CV 0.28% · H 134.22 MB

FP64 prepared-step latency

np	nh=512	nh=1024	nh=2048
1	0.08 μs CV 0.17% · H 0.004 MB	0.12 μs CV 0.21% · H 0.008 MB	0.13 μs CV 0.28% · H 0.02 MB
2	0.72 μs CV 0.34% · H 0.02 MB	1.39 μs CV 0.15% · H 0.03 MB	2.74 μs CV 0.14% · H 0.07 MB
4	1.34 μs CV 0.16% · H 0.07 MB	2.66 μs CV 0.12% · H 0.13 MB	5.25 μs CV 0.17% · H 0.26 MB
8	0.60 μs CV 0.22% · H 0.26 MB	1.02 μs CV 1.95% · H 0.52 MB	1.85 μs CV 0.13% · H 1.05 MB
16	1.45 μs CV 2.04% · H 1.05 MB	2.73 μs CV 2.69% · H 2.10 MB	5.23 μs CV 0.44% · H 4.19 MB
32	5.21 μs CV 1.22% · H 4.19 MB	10.11 μs CV 0.44% · H 8.39 MB	148.93 μs CV 2.18% · H 16.78 MB
64	121.45 μs CV 3.37% · H 16.78 MB	425.32 μs CV 0.51% · H 33.55 MB	1,590.99 μs CV 0.08% · H 67.11 MB
128	837.82 μs CV 0.53% · H 67.11 MB	2,787.57 μs CV 0.11% · H 134.22 MB	6,073.46 μs CV 0.17% · H 268.44 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

Windows 11 · NVIDIA GeForce RTX 5090 (1 GPU)

Hardware		Software and measurement	
Accelerator	NVIDIA GeForce RTX 5090	Execution environment	Local workstation
Compute resources	170 streaming multiprocessors	Operating system	Windows
Compute capability	12.0	CUDA runtime	13.4
Device memory	31.8 GiB	CUDA driver API	13.4
Device cache	96 MiB L2	Test coverage	48 cases · FP32 and FP64 · np 8-256 · nh 512-2048
Host architecture	x86-64	Numerical accuracy	48/48 passed · max abs 2.29e-13 · max rel 1.40e-05
		Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps
		Host timing control	windows single physical core (2 logical processors), mask 0xC000; normal process priority (unchanged)
		Timing views	Device-resident and host-resident
		Chart label	RTX 5090 (Windows)

Reference workload. At np=32, nh=1024, device-resident execution reaches 20.25 μ s (FP32) and 22.07 μ s (FP64); host-resident timing is shown separately to make integration cost explicit.

Timing views at np=32, nh=1024

Precision	Device-resident	Host-resident	Host-device difference (Δ / ratio)
FP32	20.25 μ s	24.08 μ s	+3.84 μ s · 1.19x
FP64	22.07 μ s	26.96 μ s	+4.89 μ s · 1.22x

Device-resident: vectors retained on the GPU. **Host-resident:** host-observed execution including vector transfers; H remains on the GPU in both views. The difference compares independent timings; it is not an isolated transfer measurement.

Windows 11 · NVIDIA GeForce RTX 5090 (1 GPU)

FP32 result matrices

FP32 device-resident latency

np	nh=512	nh=1024	nh=2048
8	20.35 μs CV 3.51% · H 0.13 MB	22.17 μs CV 0.88% · H 0.26 MB	20.83 μs CV 0.76% · H 0.52 MB
16	21.98 μs CV 0.93% · H 0.52 MB	21.76 μs CV 1.02% · H 1.05 MB	21.53 μs CV 1.15% · H 2.10 MB
32	21.59 μs CV 0.31% · H 2.10 MB	20.25 μs CV 0.47% · H 4.19 MB	21.55 μs CV 1.47% · H 8.39 MB
64	19.97 μs CV 0.78% · H 8.39 MB	21.50 μs CV 0.56% · H 16.78 MB	24.47 μs CV 0.53% · H 33.55 MB
96	21.15 μs CV 0.83% · H 18.87 MB	25.24 μs CV 0.49% · H 37.75 MB	31.62 μs CV 0.57% · H 75.50 MB
128	23.43 μs CV 0.84% · H 33.55 MB	29.83 μs CV 0.72% · H 67.11 MB	80.89 μs CV 0.36% · H 134.22 MB
192	29.79 μs CV 0.63% · H 75.50 MB	111.04 μs CV 0.26% · H 150.99 MB	200.33 μs CV 0.29% · H 301.99 MB
256	83.45 μs CV 0.59% · H 134.22 MB	176.65 μs CV 0.53% · H 268.44 MB	337.73 μs CV 0.09% · H 536.87 MB

FP32 host-resident latency

np	nh=512	nh=1024	nh=2048
8	24.64 μs CV 0.78% · H 0.13 MB	25.58 μs CV 0.76% · H 0.26 MB	24.09 μs CV 0.63% · H 0.52 MB
16	25.46 μs CV 2.88% · H 0.52 MB	25.01 μs CV 0.76% · H 1.05 MB	25.37 μs CV 1.02% · H 2.10 MB
32	25.64 μs CV 0.21% · H 2.10 MB	24.08 μs CV 0.87% · H 4.19 MB	25.53 μs CV 0.18% · H 8.39 MB
64	25.78 μs CV 0.88% · H 8.39 MB	26.34 μs CV 1.23% · H 16.78 MB	34.83 μs CV 1.77% · H 33.55 MB
96	29.24 μs CV 1.22% · H 18.87 MB	32.75 μs CV 1.01% · H 37.75 MB	39.42 μs CV 0.52% · H 75.50 MB
128	33.62 μs CV 0.93% · H 33.55 MB	39.30 μs CV 0.27% · H 67.11 MB	91.80 μs CV 0.48% · H 134.22 MB
192	48.11 μs CV 0.94% · H 75.50 MB	133.73 μs CV 0.35% · H 150.99 MB	224.45 μs CV 0.23% · H 301.99 MB
256	114.20 μs CV 0.14% · H 134.22 MB	206.01 μs CV 0.16% · H 268.44 MB	368.41 μs CV 0.05% · H 536.87 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

Windows 11 · NVIDIA GeForce RTX 5090 (1 GPU)

FP64 result matrices

FP64 device-resident latency

np	nh=512	nh=1024	nh=2048
8	20.58 μs CV 1.11% · H 0.26 MB	20.34 μs CV 1.52% · H 0.52 MB	20.86 μs CV 0.22% · H 1.05 MB
16	20.57 μs CV 0.66% · H 1.05 MB	22.00 μs CV 0.61% · H 2.10 MB	23.09 μs CV 0.26% · H 4.19 MB
32	22.45 μs CV 0.34% · H 4.19 MB	22.07 μs CV 0.30% · H 8.39 MB	37.26 μs CV 0.25% · H 16.78 MB
64	22.81 μs CV 0.77% · H 16.78 MB	25.09 μs CV 0.69% · H 33.55 MB	33.41 μs CV 0.67% · H 67.11 MB
96	25.70 μs CV 0.48% · H 37.75 MB	34.38 μs CV 0.37% · H 75.50 MB	112.75 μs CV 0.21% · H 150.99 MB
128	30.43 μs CV 0.19% · H 67.11 MB	85.16 μs CV 0.36% · H 134.22 MB	185.11 μs CV 0.17% · H 268.44 MB
192	109.93 μs CV 0.38% · H 150.99 MB	201.21 μs CV 0.48% · H 301.99 MB	382.88 μs CV 0.15% · H 603.98 MB
256	179.40 μs CV 0.23% · H 268.44 MB	336.97 μs CV 0.06% · H 536.87 MB	650.52 μs CV 0.05% · H 1,073.74 MB

FP64 host-resident latency

np	nh=512	nh=1024	nh=2048
8	22.65 μs CV 1.47% · H 0.26 MB	23.00 μs CV 1.26% · H 0.52 MB	23.98 μs CV 0.62% · H 1.05 MB
16	23.89 μs CV 1.14% · H 1.05 MB	25.35 μs CV 1.15% · H 2.10 MB	26.38 μs CV 2.37% · H 4.19 MB
32	26.61 μs CV 0.20% · H 4.19 MB	26.96 μs CV 0.27% · H 8.39 MB	41.58 μs CV 0.49% · H 16.78 MB
64	26.69 μs CV 0.31% · H 16.78 MB	30.50 μs CV 0.31% · H 33.55 MB	38.68 μs CV 0.81% · H 67.11 MB
96	31.70 μs CV 2.18% · H 37.75 MB	40.66 μs CV 1.27% · H 75.50 MB	120.46 μs CV 0.09% · H 150.99 MB
128	38.09 μs CV 0.28% · H 67.11 MB	94.21 μs CV 0.40% · H 134.22 MB	196.01 μs CV 0.14% · H 268.44 MB
192	127.98 μs CV 0.50% · H 150.99 MB	221.93 μs CV 0.32% · H 301.99 MB	407.27 μs CV 0.23% · H 603.98 MB
256	209.67 μs CV 0.36% · H 268.44 MB	364.54 μs CV 0.11% · H 536.87 MB	679.33 μs CV 0.12% · H 1,073.74 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA GeForce RTX 5090 (1 GPU)

Hardware		Software and measurement	
Accelerator	NVIDIA GeForce RTX 5090	Execution environment	RunPod
Compute resources	170 streaming multiprocessors	Operating system	Linux
Compute capability	12.0	CUDA runtime	13.3
Device memory	31.4 GiB	CUDA driver API	13.0
Device cache	96 MiB L2	Test coverage	48 cases · FP32 and FP64 · np 8-256 · nh 512-2048
Host architecture	x86-64	Numerical accuracy	48/48 passed · max abs 2.29e-13 · max rel 1.40e-05
		Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps
		Timing views	Device-resident and host-resident
		Chart label	RTX 5090

Reference workload. At np=32, nh=1024, device-resident execution reaches 14.46 μ s (FP32) and 17.28 μ s (FP64); host-resident timing is shown separately to make integration cost explicit.

Timing views at np=32, nh=1024

Precision	Device-resident	Host-resident	Host–device difference (Δ / ratio)
FP32	14.46 μ s	18.05 μ s	+3.59 μ s · 1.25x
FP64	17.28 μ s	20.44 μ s	+3.17 μ s · 1.18x

Device-resident: vectors retained on the GPU. **Host-resident:** host-observed execution including vector transfers; H remains on the GPU in both views. The difference compares independent timings; it is not an isolated transfer measurement.

RunPod · NVIDIA GeForce RTX 5090 (1 GPU)

FP32 result matrices

FP32 device-resident latency

np	nh=512	nh=1024	nh=2048
8	14.26 μs CV 0.11% · H 0.13 MB	16.37 μs CV 0.04% · H 0.26 MB	14.77 μs CV 0.08% · H 0.52 MB
16	16.44 μs CV 0.03% · H 0.52 MB	14.93 μs CV 0.01% · H 1.05 MB	17.00 μs CV 0.02% · H 2.10 MB
32	17.61 μs CV 0.19% · H 2.10 MB	14.46 μs CV 0.11% · H 4.19 MB	16.38 μs CV 0.30% · H 8.39 MB
64	15.34 μs CV 0.08% · H 8.39 MB	17.07 μs CV 0.10% · H 16.78 MB	20.33 μs CV 0.09% · H 33.55 MB
96	17.25 μs CV 0.02% · H 18.87 MB	20.92 μs CV 0.03% · H 37.75 MB	27.56 μs CV 0.07% · H 75.50 MB
128	19.29 μs CV 0.04% · H 33.55 MB	25.45 μs CV 0.03% · H 67.11 MB	77.15 μs CV 0.08% · H 134.22 MB
192	25.83 μs CV 0.02% · H 75.50 MB	107.25 μs CV 0.02% · H 150.99 MB	197.37 μs CV 0.02% · H 301.99 MB
256	79.17 μs CV 0.02% · H 134.22 MB	175.07 μs CV 0.02% · H 268.44 MB	336.88 μs CV 0.04% · H 536.87 MB

FP32 host-resident latency

np	nh=512	nh=1024	nh=2048
8	15.86 μs CV 0.12% · H 0.13 MB	18.02 μs CV 0.05% · H 0.26 MB	15.97 μs CV 0.04% · H 0.52 MB
16	17.01 μs CV 0.05% · H 0.52 MB	16.52 μs CV 0.15% · H 1.05 MB	18.35 μs CV 0.61% · H 2.10 MB
32	20.22 μs CV 0.08% · H 2.10 MB	18.05 μs CV 0.16% · H 4.19 MB	20.07 μs CV 0.05% · H 8.39 MB
64	19.10 μs CV 0.05% · H 8.39 MB	20.87 μs CV 0.08% · H 16.78 MB	24.21 μs CV 0.02% · H 33.55 MB
96	22.72 μs CV 0.14% · H 18.87 MB	26.17 μs CV 0.07% · H 37.75 MB	32.94 μs CV 0.06% · H 75.50 MB
128	26.43 μs CV 0.05% · H 33.55 MB	32.58 μs CV 0.07% · H 67.11 MB	84.02 μs CV 0.04% · H 134.22 MB
192	39.31 μs CV 0.63% · H 75.50 MB	121.68 μs CV 0.02% · H 150.99 MB	211.29 μs CV 0.01% · H 301.99 MB
256	102.76 μs CV 0.03% · H 134.22 MB	199.01 μs CV 0.03% · H 268.44 MB	360.98 μs CV 0.02% · H 536.87 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA GeForce RTX 5090 (1 GPU)

FP64 result matrices

FP64 device-resident latency

np	nh=512	nh=1024	nh=2048
8	13.51 μs CV 0.07% · H 0.26 MB	14.04 μs CV 0.03% · H 0.52 MB	14.68 μs CV 0.91% · H 1.05 MB
16	14.47 μs CV 0.05% · H 1.05 MB	16.26 μs CV 0.06% · H 2.10 MB	19.16 μs CV 0.06% · H 4.19 MB
32	17.49 μs CV 0.17% · H 4.19 MB	17.28 μs CV 0.11% · H 8.39 MB	33.07 μs CV 0.04% · H 16.78 MB
64	18.89 μs CV 0.09% · H 16.78 MB	22.69 μs CV 0.09% · H 33.55 MB	31.39 μs CV 0.03% · H 67.11 MB
96	21.35 μs CV 0.14% · H 37.75 MB	29.66 μs CV 0.04% · H 75.50 MB	109.80 μs CV 0.01% · H 150.99 MB
128	26.82 μs CV 0.07% · H 67.11 MB	79.78 μs CV 0.02% · H 134.22 MB	181.86 μs CV 0.01% · H 268.44 MB
192	105.76 μs CV 0.02% · H 150.99 MB	197.92 μs CV 0.01% · H 301.99 MB	380.28 μs CV 0.01% · H 603.98 MB
256	175.96 μs CV 0.01% · H 268.44 MB	338.95 μs CV 0.01% · H 536.87 MB	655.80 μs CV 0.08% · H 1,073.74 MB

FP64 host-resident latency

np	nh=512	nh=1024	nh=2048
8	14.70 μs CV 0.07% · H 0.26 MB	15.32 μs CV 0.03% · H 0.52 MB	16.11 μs CV 0.04% · H 1.05 MB
16	16.07 μs CV 0.04% · H 1.05 MB	17.58 μs CV 0.16% · H 2.10 MB	20.43 μs CV 0.07% · H 4.19 MB
32	20.38 μs CV 0.05% · H 4.19 MB	20.44 μs CV 0.09% · H 8.39 MB	34.91 μs CV 0.05% · H 16.78 MB
64	22.03 μs CV 0.08% · H 16.78 MB	26.29 μs CV 0.04% · H 33.55 MB	34.98 μs CV 0.04% · H 67.11 MB
96	26.46 μs CV 0.05% · H 37.75 MB	34.97 μs CV 0.05% · H 75.50 MB	114.67 μs CV 0.03% · H 150.99 MB
128	33.39 μs CV 0.47% · H 67.11 MB	86.38 μs CV 0.02% · H 134.22 MB	188.41 μs CV 0.01% · H 268.44 MB
192	118.31 μs CV 0.02% · H 150.99 MB	210.19 μs CV 0.01% · H 301.99 MB	392.52 μs CV 0.02% · H 603.98 MB
256	197.45 μs CV 0.02% · H 268.44 MB	360.60 μs CV 0.01% · H 536.87 MB	677.55 μs CV 0.03% · H 1,073.74 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA RTX PRO 6000 Blackwell (1 GPU)

Hardware		Software and measurement	
Accelerator	NVIDIA RTX PRO 6000 Blackwell Server Edition	Execution environment	RunPod
Compute resources	188 streaming multiprocessors	Operating system	Linux
Compute capability	12.0	CUDA runtime	13.3
Device memory	95.0 GiB	CUDA driver API	13.2
Device cache	128 MiB L2	Test coverage	48 cases · FP32 and FP64 · np 8-256 · nh 512-2048
Host architecture	x86-64	Numerical accuracy	48/48 passed · max abs 2.39e-13 · max rel 1.40e-05
		Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps
		Timing views	Device-resident and host-resident
		Chart label	RTX PRO 6000

Reference workload. At np=32, nh=1024, device-resident execution reaches 20.03 μ s (FP32) and 23.44 μ s (FP64); host-resident timing is shown separately to make integration cost explicit.

Timing views at np=32, nh=1024

Precision	Device-resident	Host-resident	Host-device difference (Δ / ratio)
FP32	20.03 μ s	22.50 μ s	+2.47 μ s · 1.12x
FP64	23.44 μ s	26.05 μ s	+2.61 μ s · 1.11x

Device-resident: vectors retained on the GPU. **Host-resident:** host-observed execution including vector transfers; H remains on the GPU in both views. The difference compares independent timings; it is not an isolated transfer measurement.

RunPod · NVIDIA RTX PRO 6000 Blackwell (1 GPU)

FP32 result matrices

FP32 device-resident latency

np	nh=512	nh=1024	nh=2048
8	19.79 μs CV 1.02% · H 0.13 MB	22.82 μs CV 0.06% · H 0.26 MB	19.71 μs CV 0.08% · H 0.52 MB
16	22.72 μs CV 0.12% · H 0.52 MB	20.18 μs CV 0.32% · H 1.05 MB	20.71 μs CV 0.64% · H 2.10 MB
32	20.03 μs CV 0.73% · H 2.10 MB	20.03 μs CV 0.61% · H 4.19 MB	22.21 μs CV 0.10% · H 8.39 MB
64	20.30 μs CV 0.04% · H 8.39 MB	22.32 μs CV 0.20% · H 16.78 MB	26.38 μs CV 0.43% · H 33.55 MB
96	22.02 μs CV 0.03% · H 18.87 MB	26.05 μs CV 0.05% · H 37.75 MB	34.15 μs CV 0.35% · H 75.50 MB
128	24.04 μs CV 0.12% · H 33.55 MB	32.13 μs CV 0.43% · H 67.11 MB	46.06 μs CV 0.07% · H 134.22 MB
192	30.10 μs CV 0.16% · H 75.50 MB	118.07 μs CV 0.02% · H 150.99 MB	231.43 μs CV 0.04% · H 301.99 MB
256	40.93 μs CV 0.05% · H 134.22 MB	198.83 μs CV 0.03% · H 268.44 MB	376.38 μs CV 0.01% · H 536.87 MB

FP32 host-resident latency

np	nh=512	nh=1024	nh=2048
8	19.67 μs CV 0.31% · H 0.13 MB	22.66 μs CV 0.07% · H 0.26 MB	19.37 μs CV 0.16% · H 0.52 MB
16	24.46 μs CV 0.06% · H 0.52 MB	19.62 μs CV 0.12% · H 1.05 MB	20.68 μs CV 0.33% · H 2.10 MB
32	22.20 μs CV 0.64% · H 2.10 MB	22.50 μs CV 0.61% · H 4.19 MB	24.72 μs CV 0.04% · H 8.39 MB
64	25.03 μs CV 0.05% · H 8.39 MB	27.16 μs CV 0.19% · H 16.78 MB	31.09 μs CV 0.25% · H 33.55 MB
96	29.26 μs CV 0.07% · H 18.87 MB	32.95 μs CV 0.35% · H 37.75 MB	41.07 μs CV 0.04% · H 75.50 MB
128	33.86 μs CV 0.07% · H 33.55 MB	41.27 μs CV 0.20% · H 67.11 MB	54.92 μs CV 0.03% · H 134.22 MB
192	48.34 μs CV 0.18% · H 75.50 MB	135.89 μs CV 0.02% · H 150.99 MB	249.19 μs CV 0.02% · H 301.99 MB
256	70.92 μs CV 0.05% · H 134.22 MB	229.43 μs CV 0.14% · H 268.44 MB	407.34 μs CV 0.05% · H 536.87 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA RTX PRO 6000 Blackwell (1 GPU)

FP64 result matrices

FP64 device-resident latency

np	nh=512	nh=1024	nh=2048
8	18.65 μs CV 0.90% · H 0.26 MB	19.14 μs CV 0.10% · H 0.52 MB	20.22 μs CV 0.15% · H 1.05 MB
16	19.61 μs CV 0.39% · H 1.05 MB	21.66 μs CV 0.45% · H 2.10 MB	24.68 μs CV 0.23% · H 4.19 MB
32	22.70 μs CV 1.00% · H 4.19 MB	23.44 μs CV 0.20% · H 8.39 MB	36.47 μs CV 0.06% · H 16.78 MB
64	24.58 μs CV 0.14% · H 16.78 MB	29.49 μs CV 0.48% · H 33.55 MB	40.63 μs CV 0.27% · H 67.11 MB
96	27.07 μs CV 0.04% · H 37.75 MB	39.21 μs CV 0.29% · H 75.50 MB	119.68 μs CV 0.03% · H 150.99 MB
128	33.04 μs CV 0.03% · H 67.11 MB	54.61 μs CV 0.17% · H 134.22 MB	202.38 μs CV 0.02% · H 268.44 MB
192	110.43 μs CV 0.14% · H 150.99 MB	222.11 μs CV 0.03% · H 301.99 MB	422.97 μs CV 0.02% · H 603.98 MB
256	203.42 μs CV 0.02% · H 268.44 MB	379.43 μs CV 0.01% · H 536.87 MB	731.22 μs CV 0.01% · H 1,073.74 MB

FP64 host-resident latency

np	nh=512	nh=1024	nh=2048
8	18.55 μs CV 0.45% · H 0.26 MB	19.40 μs CV 0.12% · H 0.52 MB	20.25 μs CV 0.28% · H 1.05 MB
16	19.71 μs CV 0.56% · H 1.05 MB	21.44 μs CV 0.10% · H 2.10 MB	26.63 μs CV 0.45% · H 4.19 MB
32	24.37 μs CV 0.31% · H 4.19 MB	26.05 μs CV 0.08% · H 8.39 MB	38.57 μs CV 0.05% · H 16.78 MB
64	29.25 μs CV 0.21% · H 16.78 MB	34.48 μs CV 0.07% · H 33.55 MB	45.36 μs CV 0.08% · H 67.11 MB
96	33.65 μs CV 0.08% · H 37.75 MB	45.16 μs CV 0.25% · H 75.50 MB	125.58 μs CV 0.03% · H 150.99 MB
128	41.46 μs CV 0.13% · H 67.11 MB	62.84 μs CV 0.29% · H 134.22 MB	210.38 μs CV 0.10% · H 268.44 MB
192	125.20 μs CV 0.09% · H 150.99 MB	237.45 μs CV 0.03% · H 301.99 MB	438.23 μs CV 0.04% · H 603.98 MB
256	230.12 μs CV 0.17% · H 268.44 MB	406.37 μs CV 0.01% · H 536.87 MB	758.11 μs CV 0.01% · H 1,073.74 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA A100-SXM4-80GB (1 GPU)

Hardware		Software and measurement	
Accelerator	NVIDIA A100-SXM4-80GB	Execution environment	RunPod
Compute resources	108 streaming multiprocessors	Operating system	Linux
Compute capability	8.0	CUDA runtime	13.3
Device memory	79.2 GiB	CUDA driver API	13.0
Device cache	40 MiB L2	Test coverage	48 cases · FP32 and FP64 · np 8-256 · nh 512-2048
Host architecture	x86-64	Numerical accuracy	48/48 passed · max abs 3.09e-13 · max rel 1.45e-05
		Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps
		Timing views	Device-resident and host-resident
		Chart label	A100

Reference workload. At np=32, nh=1024, device-resident execution reaches 27.66 μ s (FP32) and 27.80 μ s (FP64); host-resident timing is shown separately to make integration cost explicit.

Timing views at np=32, nh=1024

Precision	Device-resident	Host-resident	Host-device difference (Δ / ratio)
FP32	27.66 μ s	38.17 μ s	+10.51 μ s · 1.38x
FP64	27.80 μ s	38.45 μ s	+10.65 μ s · 1.38x

Device-resident: vectors retained on the GPU. **Host-resident:** host-observed execution including vector transfers; H remains on the GPU in both views. The difference compares independent timings; it is not an isolated transfer measurement.

RunPod · NVIDIA A100-SXM4-80GB (1 GPU)

FP32 result matrices

FP32 device-resident latency

np	nh=512	nh=1024	nh=2048
8	27.01 μs CV 0.82% · H 0.13 MB	28.24 μs CV 0.17% · H 0.26 MB	29.43 μs CV 0.66% · H 0.52 MB
16	28.60 μs CV 0.66% · H 0.52 MB	30.33 μs CV 0.83% · H 1.05 MB	31.65 μs CV 4.60% · H 2.10 MB
32	27.06 μs CV 0.94% · H 2.10 MB	27.66 μs CV 1.78% · H 4.19 MB	30.83 μs CV 0.17% · H 8.39 MB
64	28.05 μs CV 1.46% · H 8.39 MB	32.57 μs CV 0.83% · H 16.78 MB	43.90 μs CV 0.56% · H 33.55 MB
96	30.15 μs CV 0.41% · H 18.87 MB	45.26 μs CV 0.19% · H 37.75 MB	68.67 μs CV 0.31% · H 75.50 MB
128	46.83 μs CV 0.48% · H 33.55 MB	64.51 μs CV 0.57% · H 67.11 MB	107.57 μs CV 0.37% · H 134.22 MB
192	68.40 μs CV 0.13% · H 75.50 MB	119.59 μs CV 0.96% · H 150.99 MB	230.35 μs CV 0.61% · H 301.99 MB
256	109.39 μs CV 0.58% · H 134.22 MB	190.54 μs CV 0.23% · H 268.44 MB	346.79 μs CV 0.33% · H 536.87 MB

FP32 host-resident latency

np	nh=512	nh=1024	nh=2048
8	33.91 μs CV 4.69% · H 0.13 MB	34.98 μs CV 0.47% · H 0.26 MB	35.86 μs CV 0.66% · H 0.52 MB
16	32.67 μs CV 0.37% · H 0.52 MB	37.38 μs CV 0.90% · H 1.05 MB	38.29 μs CV 1.96% · H 2.10 MB
32	37.04 μs CV 0.45% · H 2.10 MB	38.17 μs CV 0.69% · H 4.19 MB	41.47 μs CV 0.80% · H 8.39 MB
64	39.96 μs CV 2.70% · H 8.39 MB	44.97 μs CV 1.04% · H 16.78 MB	56.51 μs CV 0.22% · H 33.55 MB
96	45.65 μs CV 0.79% · H 18.87 MB	62.49 μs CV 0.38% · H 37.75 MB	86.15 μs CV 0.14% · H 75.50 MB
128	71.25 μs CV 1.97% · H 33.55 MB	87.91 μs CV 0.49% · H 67.11 MB	131.41 μs CV 0.74% · H 134.22 MB
192	109.50 μs CV 0.76% · H 75.50 MB	160.68 μs CV 0.59% · H 150.99 MB	266.77 μs CV 0.47% · H 301.99 MB
256	178.17 μs CV 1.19% · H 134.22 MB	251.96 μs CV 0.91% · H 268.44 MB	404.56 μs CV 0.19% · H 536.87 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA A100-SXM4-80GB (1 GPU)

FP64 result matrices

FP64 device-resident latency

np	nh=512	nh=1024	nh=2048
8	25.85 μs CV 0.38% · H 0.26 MB	26.79 μs CV 0.17% · H 0.52 MB	27.97 μs CV 0.53% · H 1.05 MB
16	27.71 μs CV 0.52% · H 1.05 MB	30.39 μs CV 0.52% · H 2.10 MB	30.75 μs CV 1.23% · H 4.19 MB
32	25.72 μs CV 0.70% · H 4.19 MB	27.80 μs CV 0.49% · H 8.39 MB	31.16 μs CV 0.90% · H 16.78 MB
64	28.43 μs CV 1.17% · H 16.78 MB	38.71 μs CV 0.21% · H 33.55 MB	62.25 μs CV 0.15% · H 67.11 MB
96	41.32 μs CV 2.39% · H 37.75 MB	61.56 μs CV 0.16% · H 75.50 MB	113.78 μs CV 1.42% · H 150.99 MB
128	59.65 μs CV 0.54% · H 67.11 MB	94.69 μs CV 1.07% · H 134.22 MB	177.19 μs CV 0.42% · H 268.44 MB
192	111.15 μs CV 0.22% · H 150.99 MB	200.72 μs CV 0.73% · H 301.99 MB	373.46 μs CV 0.17% · H 603.98 MB
256	176.39 μs CV 0.09% · H 268.44 MB	331.99 μs CV 0.60% · H 536.87 MB	638.56 μs CV 0.51% · H 1,073.74 MB

FP64 host-resident latency

np	nh=512	nh=1024	nh=2048
8	31.74 μs CV 0.67% · H 0.26 MB	33.13 μs CV 0.97% · H 0.52 MB	34.20 μs CV 1.74% · H 1.05 MB
16	33.93 μs CV 0.83% · H 1.05 MB	36.87 μs CV 1.13% · H 2.10 MB	37.29 μs CV 0.83% · H 4.19 MB
32	35.75 μs CV 0.89% · H 4.19 MB	38.45 μs CV 0.65% · H 8.39 MB	41.09 μs CV 0.45% · H 16.78 MB
64	40.49 μs CV 0.65% · H 16.78 MB	52.13 μs CV 0.80% · H 33.55 MB	74.43 μs CV 0.20% · H 67.11 MB
96	57.70 μs CV 0.48% · H 37.75 MB	78.12 μs CV 0.27% · H 75.50 MB	128.48 μs CV 0.63% · H 150.99 MB
128	82.63 μs CV 0.19% · H 67.11 MB	116.90 μs CV 1.45% · H 134.22 MB	199.27 μs CV 1.08% · H 268.44 MB
192	150.34 μs CV 0.73% · H 150.99 MB	233.72 μs CV 0.73% · H 301.99 MB	412.16 μs CV 0.92% · H 603.98 MB
256	233.83 μs CV 0.11% · H 268.44 MB	391.93 μs CV 0.84% · H 536.87 MB	701.69 μs CV 0.46% · H 1,073.74 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA H100 80GB HBM3 (1 GPU)

Hardware		Software and measurement	
Accelerator	NVIDIA H100 80GB HBM3	Execution environment	RunPod
Compute resources	132 streaming multiprocessors	Operating system	Linux
Compute capability	9.0	CUDA runtime	13.3
Device memory	79.2 GiB	CUDA driver API	13.0
Device cache	50 MiB L2	Test coverage	48 cases · FP32 and FP64 · np 8-256 · nh 512-2048
Host architecture	x86-64	Numerical accuracy	48/48 passed · max abs 3.69e-13 · max rel 1.01e-05
		Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps
		Timing views	Device-resident and host-resident
		Chart label	H100

Reference workload. At np=32, nh=1024, device-resident execution reaches 21.52 μ s (FP32) and 21.53 μ s (FP64); host-resident timing is shown separately to make integration cost explicit.

Timing views at np=32, nh=1024

Precision	Device-resident	Host-resident	Host-device difference (Δ / ratio)
FP32	21.52 μ s	28.34 μ s	+6.82 μ s · 1.32x
FP64	21.53 μ s	27.02 μ s	+5.50 μ s · 1.26x

Device-resident: vectors retained on the GPU. **Host-resident:** host-observed execution including vector transfers; H remains on the GPU in both views. The difference compares independent timings; it is not an isolated transfer measurement.

RunPod · NVIDIA H100 80GB HBM3 (1 GPU)

FP32 result matrices

FP32 device-resident latency

np	nh=512	nh=1024	nh=2048
8	20.75 μs CV 0.03% · H 0.13 MB	21.62 μs CV 1.18% · H 0.26 MB	22.20 μs CV 1.41% · H 0.52 MB
16	21.57 μs CV 0.03% · H 0.52 MB	22.53 μs CV 0.05% · H 1.05 MB	23.68 μs CV 0.02% · H 2.10 MB
32	20.82 μs CV 0.11% · H 2.10 MB	21.52 μs CV 0.04% · H 4.19 MB	23.23 μs CV 1.29% · H 8.39 MB
64	21.69 μs CV 0.02% · H 8.39 MB	22.72 μs CV 0.11% · H 16.78 MB	26.92 μs CV 0.02% · H 33.55 MB
96	22.66 μs CV 0.06% · H 18.87 MB	38.05 μs CV 0.16% · H 37.75 MB	48.26 μs CV 0.02% · H 75.50 MB
128	25.76 μs CV 0.03% · H 33.55 MB	42.11 μs CV 0.03% · H 67.11 MB	67.60 μs CV 0.02% · H 134.22 MB
192	44.05 μs CV 0.03% · H 75.50 MB	75.46 μs CV 0.04% · H 150.99 MB	128.92 μs CV 0.02% · H 301.99 MB
256	65.36 μs CV 0.02% · H 134.22 MB	112.13 μs CV 0.05% · H 268.44 MB	206.07 μs CV 0.21% · H 536.87 MB

FP32 host-resident latency

np	nh=512	nh=1024	nh=2048
8	26.37 μs CV 0.45% · H 0.13 MB	24.38 μs CV 0.09% · H 0.26 MB	25.58 μs CV 0.08% · H 0.52 MB
16	24.57 μs CV 0.07% · H 0.52 MB	25.76 μs CV 0.06% · H 1.05 MB	26.83 μs CV 0.03% · H 2.10 MB
32	27.16 μs CV 0.05% · H 2.10 MB	28.34 μs CV 0.11% · H 4.19 MB	29.42 μs CV 0.26% · H 8.39 MB
64	30.78 μs CV 0.04% · H 8.39 MB	29.93 μs CV 0.15% · H 16.78 MB	36.09 μs CV 0.03% · H 33.55 MB
96	33.20 μs CV 0.07% · H 18.87 MB	50.75 μs CV 0.10% · H 37.75 MB	60.79 μs CV 0.03% · H 75.50 MB
128	42.45 μs CV 0.05% · H 33.55 MB	58.79 μs CV 0.04% · H 67.11 MB	82.89 μs CV 0.07% · H 134.22 MB
192	72.87 μs CV 0.04% · H 75.50 MB	103.74 μs CV 0.04% · H 150.99 MB	157.16 μs CV 0.01% · H 301.99 MB
256	112.52 μs CV 0.14% · H 134.22 MB	159.16 μs CV 0.86% · H 268.44 MB	252.52 μs CV 0.23% · H 536.87 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA H100 80GB HBM3 (1 GPU)

FP64 result matrices

FP64 device-resident latency

np	nh=512	nh=1024	nh=2048
8	21.38 μs CV 0.58% · H 0.26 MB	19.85 μs CV 0.05% · H 0.52 MB	21.05 μs CV 0.69% · H 1.05 MB
16	20.45 μs CV 0.04% · H 1.05 MB	21.96 μs CV 0.04% · H 2.10 MB	24.78 μs CV 0.06% · H 4.19 MB
32	21.53 μs CV 0.02% · H 4.19 MB	21.53 μs CV 0.16% · H 8.39 MB	23.45 μs CV 0.03% · H 16.78 MB
64	21.26 μs CV 0.04% · H 16.78 MB	25.37 μs CV 0.13% · H 33.55 MB	41.86 μs CV 0.07% · H 67.11 MB
96	26.54 μs CV 0.07% · H 37.75 MB	44.69 μs CV 0.03% · H 75.50 MB	68.39 μs CV 0.01% · H 150.99 MB
128	40.63 μs CV 0.03% · H 67.11 MB	62.39 μs CV 0.03% · H 134.22 MB	106.55 μs CV 0.02% · H 268.44 MB
192	70.27 μs CV 0.02% · H 150.99 MB	119.73 μs CV 0.38% · H 301.99 MB	216.74 μs CV 0.02% · H 603.98 MB
256	106.12 μs CV 0.02% · H 268.44 MB	193.06 μs CV 0.35% · H 536.87 MB	366.92 μs CV 0.50% · H 1,073.74 MB

FP64 host-resident latency

np	nh=512	nh=1024	nh=2048
8	25.00 μs CV 4.09% · H 0.26 MB	22.72 μs CV 0.06% · H 0.52 MB	24.12 μs CV 0.07% · H 1.05 MB
16	23.47 μs CV 0.08% · H 1.05 MB	24.94 μs CV 1.48% · H 2.10 MB	28.10 μs CV 0.05% · H 4.19 MB
32	26.67 μs CV 0.03% · H 4.19 MB	27.02 μs CV 0.04% · H 8.39 MB	29.75 μs CV 1.55% · H 16.78 MB
64	27.97 μs CV 0.02% · H 16.78 MB	33.78 μs CV 0.10% · H 33.55 MB	50.08 μs CV 0.09% · H 67.11 MB
96	37.85 μs CV 0.15% · H 37.75 MB	56.05 μs CV 0.06% · H 75.50 MB	79.90 μs CV 0.03% · H 150.99 MB
128	55.44 μs CV 0.06% · H 67.11 MB	78.01 μs CV 0.03% · H 134.22 MB	120.76 μs CV 0.02% · H 268.44 MB
192	95.24 μs CV 0.20% · H 150.99 MB	142.31 μs CV 0.32% · H 301.99 MB	241.66 μs CV 0.02% · H 603.98 MB
256	148.82 μs CV 0.16% · H 268.44 MB	237.33 μs CV 0.10% · H 536.87 MB	411.41 μs CV 0.36% · H 1,073.74 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA H200 (1 GPU)

Hardware		Software and measurement	
Accelerator	NVIDIA H200	Execution environment	RunPod
Compute resources	132 streaming multiprocessors	Operating system	Linux
Compute capability	9.0	CUDA runtime	13.3
Device memory	139.8 GiB	CUDA driver API	13.0
Device cache	60 MiB L2	Test coverage	48 cases · FP32 and FP64 · np 8-256 · nh 512-2048
Host architecture	x86-64	Numerical accuracy	48/48 passed · max abs 3.69e-13 · max rel 1.01e-05
		Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps
		Timing views	Device-resident and host-resident
		Chart label	H200

Reference workload. At np=32, nh=1024, device-resident execution reaches 20.90 μ s (FP32) and 20.24 μ s (FP64); host-resident timing is shown separately to make integration cost explicit.

Timing views at np=32, nh=1024

Precision	Device-resident	Host-resident	Host-device difference (Δ / ratio)
FP32	20.90 μ s	25.54 μ s	+4.64 μ s · 1.22x
FP64	20.24 μ s	25.67 μ s	+5.43 μ s · 1.27x

Device-resident: vectors retained on the GPU. **Host-resident:** host-observed execution including vector transfers; H remains on the GPU in both views. The difference compares independent timings; it is not an isolated transfer measurement.

RunPod · NVIDIA H200 (1 GPU)

FP32 result matrices

FP32 device-resident latency

np	nh=512	nh=1024	nh=2048
8	19.52 μs CV 0.05% · H 0.13 MB	20.42 μs CV 0.06% · H 0.26 MB	20.93 μs CV 0.03% · H 0.52 MB
16	20.33 μs CV 0.02% · H 0.52 MB	20.96 μs CV 0.03% · H 1.05 MB	22.73 μs CV 0.02% · H 2.10 MB
32	19.57 μs CV 0.05% · H 2.10 MB	20.90 μs CV 0.03% · H 4.19 MB	22.56 μs CV 0.07% · H 8.39 MB
64	20.41 μs CV 0.05% · H 8.39 MB	21.70 μs CV 0.16% · H 16.78 MB	25.09 μs CV 0.07% · H 33.55 MB
96	21.43 μs CV 0.20% · H 18.87 MB	32.12 μs CV 0.29% · H 37.75 MB	42.30 μs CV 0.03% · H 75.50 MB
128	24.28 μs CV 0.03% · H 33.55 MB	37.58 μs CV 0.02% · H 67.11 MB	56.91 μs CV 0.02% · H 134.22 MB
192	38.83 μs CV 0.01% · H 75.50 MB	59.93 μs CV 0.04% · H 150.99 MB	96.58 μs CV 0.02% · H 301.99 MB
256	55.49 μs CV 1.14% · H 134.22 MB	83.29 μs CV 0.03% · H 268.44 MB	145.18 μs CV 0.47% · H 536.87 MB

FP32 host-resident latency

np	nh=512	nh=1024	nh=2048
8	21.94 μs CV 0.05% · H 0.13 MB	22.46 μs CV 0.29% · H 0.26 MB	23.52 μs CV 0.07% · H 0.52 MB
16	22.40 μs CV 0.02% · H 0.52 MB	23.23 μs CV 0.07% · H 1.05 MB	24.95 μs CV 0.04% · H 2.10 MB
32	24.51 μs CV 0.05% · H 2.10 MB	25.54 μs CV 0.13% · H 4.19 MB	27.78 μs CV 0.05% · H 8.39 MB
64	27.38 μs CV 0.16% · H 8.39 MB	29.05 μs CV 0.19% · H 16.78 MB	32.62 μs CV 0.04% · H 33.55 MB
96	31.08 μs CV 0.02% · H 18.87 MB	41.76 μs CV 0.41% · H 37.75 MB	52.71 μs CV 0.02% · H 75.50 MB
128	37.27 μs CV 0.06% · H 33.55 MB	50.92 μs CV 0.08% · H 67.11 MB	70.41 μs CV 0.06% · H 134.22 MB
192	62.06 μs CV 0.35% · H 75.50 MB	82.04 μs CV 0.30% · H 150.99 MB	118.87 μs CV 0.05% · H 301.99 MB
256	91.98 μs CV 0.38% · H 134.22 MB	120.12 μs CV 0.04% · H 268.44 MB	182.64 μs CV 0.47% · H 536.87 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA H200 (1 GPU)

FP64 result matrices

FP64 device-resident latency

np	nh=512	nh=1024	nh=2048
8	18.62 μs CV 0.02% · H 0.26 MB	18.82 μs CV 0.12% · H 0.52 MB	20.02 μs CV 0.27% · H 1.05 MB
16	19.53 μs CV 0.03% · H 1.05 MB	20.50 μs CV 0.05% · H 2.10 MB	23.78 μs CV 0.02% · H 4.19 MB
32	19.94 μs CV 0.13% · H 4.19 MB	20.24 μs CV 0.02% · H 8.39 MB	22.79 μs CV 0.09% · H 16.78 MB
64	19.96 μs CV 0.03% · H 16.78 MB	24.10 μs CV 0.02% · H 33.55 MB	39.18 μs CV 0.05% · H 67.11 MB
96	23.47 μs CV 0.02% · H 37.75 MB	39.80 μs CV 0.05% · H 75.50 MB	52.53 μs CV 0.30% · H 150.99 MB
128	37.45 μs CV 0.02% · H 67.11 MB	53.69 μs CV 0.04% · H 134.22 MB	79.20 μs CV 0.02% · H 268.44 MB
192	55.30 μs CV 0.01% · H 150.99 MB	86.64 μs CV 0.05% · H 301.99 MB	155.71 μs CV 0.01% · H 603.98 MB
256	79.46 μs CV 0.05% · H 268.44 MB	140.38 μs CV 0.02% · H 536.87 MB	256.25 μs CV 0.25% · H 1,073.74 MB

FP64 host-resident latency

np	nh=512	nh=1024	nh=2048
8	20.98 μs CV 0.30% · H 0.26 MB	21.31 μs CV 0.02% · H 0.52 MB	22.55 μs CV 0.04% · H 1.05 MB
16	21.95 μs CV 0.06% · H 1.05 MB	22.84 μs CV 0.03% · H 2.10 MB	25.91 μs CV 0.04% · H 4.19 MB
32	24.90 μs CV 0.03% · H 4.19 MB	25.67 μs CV 0.11% · H 8.39 MB	27.36 μs CV 0.06% · H 16.78 MB
64	27.18 μs CV 0.04% · H 16.78 MB	30.59 μs CV 0.13% · H 33.55 MB	47.19 μs CV 0.03% · H 67.11 MB
96	32.92 μs CV 0.03% · H 37.75 MB	50.28 μs CV 0.08% · H 75.50 MB	62.47 μs CV 0.04% · H 150.99 MB
128	49.25 μs CV 0.14% · H 67.11 MB	66.32 μs CV 0.02% · H 134.22 MB	90.30 μs CV 0.06% · H 268.44 MB
192	75.16 μs CV 0.25% · H 150.99 MB	107.25 μs CV 0.34% · H 301.99 MB	175.17 μs CV 0.02% · H 603.98 MB
256	113.29 μs CV 0.07% · H 268.44 MB	173.98 μs CV 2.55% · H 536.87 MB	290.32 μs CV 0.15% · H 1,073.74 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA B200 (1 GPU)

Hardware		Software and measurement	
Accelerator	NVIDIA B200	Execution environment	RunPod
Compute resources	148 streaming multiprocessors	Operating system	Linux
Compute capability	10.0	CUDA runtime	13.3
Device memory	178.4 GiB	CUDA driver API	13.0
Device cache	126 MiB L2	Test coverage	48 cases · FP32 and FP64 · np 8-256 · nh 512-2048
Host architecture	x86-64	Numerical accuracy	48/48 passed · max abs 3.44e-13 · max rel 3.16e-05
		Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps
		Timing views	Device-resident and host-resident
		Chart label	B200

Reference workload. At np=32, nh=1024, device-resident execution reaches 23.85 μ s (FP32) and 21.94 μ s (FP64); host-resident timing is shown separately to make integration cost explicit.

Timing views at np=32, nh=1024

Precision	Device-resident	Host-resident	Host-device difference (Δ / ratio)
FP32	23.85 μ s	27.71 μ s	+3.86 μ s · 1.16x
FP64	21.94 μ s	27.74 μ s	+5.79 μ s · 1.26x

Device-resident: vectors retained on the GPU. **Host-resident:** host-observed execution including vector transfers; H remains on the GPU in both views. The difference compares independent timings; it is not an isolated transfer measurement.

RunPod · NVIDIA B200 (1 GPU)

FP32 result matrices

FP32 device-resident latency

np	nh=512	nh=1024	nh=2048
8	21.08 μs CV 1.01% · H 0.13 MB	23.03 μs CV 0.27% · H 0.26 MB	23.67 μs CV 0.09% · H 0.52 MB
16	22.61 μs CV 0.14% · H 0.52 MB	22.90 μs CV 0.03% · H 1.05 MB	24.91 μs CV 0.10% · H 2.10 MB
32	22.93 μs CV 0.20% · H 2.10 MB	23.85 μs CV 0.17% · H 4.19 MB	23.09 μs CV 0.26% · H 8.39 MB
64	21.56 μs CV 0.05% · H 8.39 MB	22.97 μs CV 0.07% · H 16.78 MB	30.98 μs CV 0.05% · H 33.55 MB
96	24.42 μs CV 0.07% · H 18.87 MB	32.83 μs CV 0.11% · H 37.75 MB	43.30 μs CV 0.04% · H 75.50 MB
128	26.81 μs CV 0.12% · H 33.55 MB	38.11 μs CV 0.07% · H 67.11 MB	64.63 μs CV 0.03% · H 134.22 MB
192	38.00 μs CV 0.13% · H 75.50 MB	64.49 μs CV 0.06% · H 150.99 MB	97.97 μs CV 0.03% · H 301.99 MB
256	63.57 μs CV 0.03% · H 134.22 MB	98.66 μs CV 0.01% · H 268.44 MB	162.32 μs CV 0.03% · H 536.87 MB

FP32 host-resident latency

np	nh=512	nh=1024	nh=2048
8	24.12 μs CV 0.07% · H 0.13 MB	24.50 μs CV 0.14% · H 0.26 MB	25.24 μs CV 0.08% · H 0.52 MB
16	25.50 μs CV 0.05% · H 0.52 MB	24.93 μs CV 0.02% · H 1.05 MB	26.86 μs CV 0.06% · H 2.10 MB
32	26.42 μs CV 0.13% · H 2.10 MB	27.71 μs CV 0.05% · H 4.19 MB	27.63 μs CV 0.06% · H 8.39 MB
64	28.42 μs CV 0.01% · H 8.39 MB	29.19 μs CV 0.20% · H 16.78 MB	37.03 μs CV 0.09% · H 33.55 MB
96	31.81 μs CV 0.12% · H 18.87 MB	40.25 μs CV 0.01% · H 37.75 MB	51.13 μs CV 0.03% · H 75.50 MB
128	37.33 μs CV 0.16% · H 33.55 MB	48.28 μs CV 0.05% · H 67.11 MB	74.10 μs CV 0.01% · H 134.22 MB
192	57.01 μs CV 0.03% · H 75.50 MB	82.37 μs CV 0.03% · H 150.99 MB	116.82 μs CV 0.01% · H 301.99 MB
256	95.23 μs CV 0.10% · H 134.22 MB	129.73 μs CV 0.01% · H 268.44 MB	193.64 μs CV 0.01% · H 536.87 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA B200 (1 GPU)

FP64 result matrices

FP64 device-resident latency

np	nh=512	nh=1024	nh=2048
8	21.77 μs CV 0.96% · H 0.26 MB	21.20 μs CV 0.23% · H 0.52 MB	22.01 μs CV 0.12% · H 1.05 MB
16	22.20 μs CV 0.10% · H 1.05 MB	22.21 μs CV 0.09% · H 2.10 MB	24.22 μs CV 0.14% · H 4.19 MB
32	21.95 μs CV 0.02% · H 4.19 MB	21.94 μs CV 0.05% · H 8.39 MB	24.00 μs CV 0.06% · H 16.78 MB
64	21.93 μs CV 0.04% · H 16.78 MB	24.55 μs CV 0.26% · H 33.55 MB	26.33 μs CV 0.05% · H 67.11 MB
96	24.09 μs CV 0.15% · H 37.75 MB	27.84 μs CV 1.13% · H 75.50 MB	45.89 μs CV 0.09% · H 150.99 MB
128	24.60 μs CV 0.13% · H 67.11 MB	42.18 μs CV 0.09% · H 134.22 MB	65.47 μs CV 0.03% · H 268.44 MB
192	48.74 μs CV 0.05% · H 150.99 MB	71.58 μs CV 0.05% · H 301.99 MB	115.75 μs CV 0.36% · H 603.98 MB
256	64.06 μs CV 0.05% · H 268.44 MB	102.12 μs CV 0.02% · H 536.87 MB	176.51 μs CV 0.02% · H 1,073.74 MB

FP64 host-resident latency

np	nh=512	nh=1024	nh=2048
8	24.99 μs CV 0.12% · H 0.26 MB	24.24 μs CV 0.70% · H 0.52 MB	25.57 μs CV 0.09% · H 1.05 MB
16	25.17 μs CV 0.04% · H 1.05 MB	26.23 μs CV 0.05% · H 2.10 MB	28.65 μs CV 0.16% · H 4.19 MB
32	26.56 μs CV 0.02% · H 4.19 MB	27.74 μs CV 0.02% · H 8.39 MB	28.79 μs CV 0.28% · H 16.78 MB
64	27.91 μs CV 0.11% · H 16.78 MB	31.25 μs CV 0.05% · H 33.55 MB	32.46 μs CV 0.03% · H 67.11 MB
96	32.40 μs CV 0.03% · H 37.75 MB	36.30 μs CV 0.09% · H 75.50 MB	52.65 μs CV 0.05% · H 150.99 MB
128	33.97 μs CV 0.04% · H 67.11 MB	51.99 μs CV 0.04% · H 134.22 MB	75.71 μs CV 0.02% · H 268.44 MB
192	64.45 μs CV 0.05% · H 150.99 MB	88.62 μs CV 0.04% · H 301.99 MB	131.70 μs CV 0.01% · H 603.98 MB
256	91.85 μs CV 0.07% · H 268.44 MB	130.67 μs CV 0.05% · H 536.87 MB	203.82 μs CV 0.03% · H 1,073.74 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA RTX PRO 6000 Blackwell Workstation Edition (4 GPUs)

Hardware		Software and measurement	
Accelerator	4 × NVIDIA RTX PRO 6000 Blackwell Workstation Edition	Execution environment	RunPod
Compute resources	4 GPUs; 188 streaming multiprocessors per GPU	Operating system	Linux
Compute capability	12.0	CUDA runtime	13.3
Device memory	4 × 95.0 GiB	CUDA driver API	13.0
Device cache	4 × 128 MiB L2	Test coverage	192 cases across 4 GPU configurations · FP32 and FP64 · np 8-256 · nh 512-2048
Host architecture	x86-64	Numerical accuracy	192/192 passed · max abs 4.43e-13 · max rel 1.40e-05
		Measurement	7 repeats · 250 ms target per repeat · 64 warm-up steps
		Timing views	Device-resident and host-resident
		Chart label	RTX PRO 6000 Workstation × 4

Large-operator scaling case. At np=256, nh=2048, the measured scaling configurations show: FP32 peaks at 4 GPUs (4.19x), while FP64 reaches 2.93x with 4 GPUs.

Measured GPU configurations at np=32, nh=1024

GPUs	Precision	Device-resident	Host-resident	Host-device difference (Δ / ratio)
1	FP32	16.52 μ s	23.09 μ s	+6.57 μ s · 1.40x
1	FP64	19.94 μ s	25.90 μ s	+5.96 μ s · 1.30x
2	FP32	26.24 μ s	31.63 μ s	+5.39 μ s · 1.21x
2	FP64	40.46 μ s	33.88 μ s	-6.58 μ s · 0.84x
3	FP32	54.06 μ s	40.14 μ s	-13.93 μ s · 0.74x
3	FP64	52.89 μ s	40.48 μ s	-12.41 μ s · 0.77x
4	FP32	57.03 μ s	48.36 μ s	-8.67 μ s · 0.85x
4	FP64	59.12 μ s	48.43 μ s	-10.69 μ s · 0.82x

Each GPU-count configuration includes a complete 48-case measurement. The following subsections provide both residency matrices for every configuration. Δ and ratio compare independent execution paths, not isolated transfer cost; a negative Δ can occur and does not imply negative transfer time.

Four measured GPU configurations. The scaling study covers 1, 2, 3, and 4 GPUs. Reported latency is the complete prepared step across the GPUs in each configuration.

RunPod · NVIDIA RTX PRO 6000 Blackwell Workstation Edition (4 GPUs)

1-GPU configuration · FP32 result matrices

FP32 device-resident latency

np	nh=512	nh=1024	nh=2048
8	16.81 μs CV 0.24% · H 0.13 MB	19.32 μs CV 0.65% · H 0.26 MB	17.34 μs CV 0.10% · H 0.52 MB
16	19.16 μs CV 0.49% · H 0.52 MB	17.46 μs CV 0.52% · H 1.05 MB	17.96 μs CV 0.03% · H 2.10 MB
32	16.56 μs CV 0.09% · H 2.10 MB	16.52 μs CV 0.15% · H 4.19 MB	18.45 μs CV 0.44% · H 8.39 MB
64	17.49 μs CV 0.34% · H 8.39 MB	19.25 μs CV 0.12% · H 16.78 MB	22.05 μs CV 0.07% · H 33.55 MB
96	19.63 μs CV 0.54% · H 18.87 MB	21.57 μs CV 0.10% · H 37.75 MB	27.97 μs CV 0.22% · H 75.50 MB
128	21.06 μs CV 0.46% · H 33.55 MB	27.36 μs CV 0.15% · H 67.11 MB	40.80 μs CV 0.07% · H 134.22 MB
192	25.69 μs CV 0.19% · H 75.50 MB	107.52 μs CV 0.12% · H 150.99 MB	210.26 μs CV 0.04% · H 301.99 MB
256	35.11 μs CV 0.12% · H 134.22 MB	182.57 μs CV 0.06% · H 268.44 MB	347.65 μs CV 0.03% · H 536.87 MB

FP32 host-resident latency

np	nh=512	nh=1024	nh=2048
8	21.01 μs CV 0.17% · H 0.13 MB	23.50 μs CV 0.42% · H 0.26 MB	21.24 μs CV 0.20% · H 0.52 MB
16	23.13 μs CV 0.10% · H 0.52 MB	21.03 μs CV 0.06% · H 1.05 MB	21.93 μs CV 0.05% · H 2.10 MB
32	22.97 μs CV 0.11% · H 2.10 MB	23.09 μs CV 0.13% · H 4.19 MB	24.98 μs CV 0.10% · H 8.39 MB
64	28.34 μs CV 0.13% · H 8.39 MB	30.18 μs CV 0.19% · H 16.78 MB	30.81 μs CV 0.13% · H 33.55 MB
96	33.97 μs CV 0.21% · H 18.87 MB	34.39 μs CV 0.18% · H 37.75 MB	43.37 μs CV 0.07% · H 75.50 MB
128	38.79 μs CV 0.17% · H 33.55 MB	47.88 μs CV 0.06% · H 67.11 MB	61.56 μs CV 0.25% · H 134.22 MB
192	64.60 μs CV 2.29% · H 75.50 MB	146.16 μs CV 0.23% · H 150.99 MB	248.46 μs CV 0.04% · H 301.99 MB
256	98.84 μs CV 0.05% · H 134.22 MB	246.64 μs CV 0.06% · H 268.44 MB	409.41 μs CV 0.12% · H 536.87 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA RTX PRO 6000 Blackwell Workstation Edition (4 GPUs)

1-GPU configuration · FP64 result matrices

FP64 device-resident latency

np	nh=512	nh=1024	nh=2048
8	15.91 μs CV 0.17% · H 0.26 MB	16.34 μs CV 0.42% · H 0.52 MB	17.22 μs CV 0.19% · H 1.05 MB
16	16.77 μs CV 0.14% · H 1.05 MB	18.39 μs CV 0.04% · H 2.10 MB	20.45 μs CV 0.62% · H 4.19 MB
32	18.57 μs CV 0.09% · H 4.19 MB	19.94 μs CV 0.06% · H 8.39 MB	29.43 μs CV 0.12% · H 16.78 MB
64	20.49 μs CV 0.04% · H 16.78 MB	24.63 μs CV 0.03% · H 33.55 MB	33.19 μs CV 0.06% · H 67.11 MB
96	22.45 μs CV 0.07% · H 37.75 MB	32.40 μs CV 0.25% · H 75.50 MB	109.72 μs CV 0.04% · H 150.99 MB
128	26.75 μs CV 0.13% · H 67.11 MB	45.92 μs CV 0.13% · H 134.22 MB	186.45 μs CV 0.05% · H 268.44 MB
192	96.75 μs CV 0.11% · H 150.99 MB	206.01 μs CV 0.07% · H 301.99 MB	392.80 μs CV 0.03% · H 603.98 MB
256	187.46 μs CV 0.01% · H 268.44 MB	351.16 μs CV 0.02% · H 536.87 MB	679.73 μs CV 0.01% · H 1,073.74 MB

FP64 host-resident latency

np	nh=512	nh=1024	nh=2048
8	19.24 μs CV 0.15% · H 0.26 MB	19.88 μs CV 0.05% · H 0.52 MB	20.74 μs CV 0.07% · H 1.05 MB
16	20.69 μs CV 0.09% · H 1.05 MB	21.94 μs CV 0.20% · H 2.10 MB	24.23 μs CV 0.25% · H 4.19 MB
32	24.52 μs CV 0.12% · H 4.19 MB	25.90 μs CV 0.18% · H 8.39 MB	35.17 μs CV 0.11% · H 16.78 MB
64	30.67 μs CV 0.14% · H 16.78 MB	35.00 μs CV 0.17% · H 33.55 MB	41.58 μs CV 0.18% · H 67.11 MB
96	36.67 μs CV 0.32% · H 37.75 MB	46.42 μs CV 0.03% · H 75.50 MB	123.67 μs CV 0.22% · H 150.99 MB
128	46.85 μs CV 0.47% · H 67.11 MB	65.84 μs CV 0.06% · H 134.22 MB	206.24 μs CV 0.08% · H 268.44 MB
192	134.90 μs CV 0.35% · H 150.99 MB	243.70 μs CV 0.22% · H 301.99 MB	429.76 μs CV 0.02% · H 603.98 MB
256	250.90 μs CV 0.12% · H 268.44 MB	414.14 μs CV 0.09% · H 536.87 MB	740.80 μs CV 0.10% · H 1,073.74 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA RTX PRO 6000 Blackwell Workstation Edition (4 GPUs)

2-GPU configuration · FP32 result matrices

FP32 device-resident latency

np	nh=512	nh=1024	nh=2048
8	25.16 μs CV 0.40% · H 0.13 MB	28.90 μs CV 0.57% · H 0.26 MB	42.84 μs CV 0.17% · H 0.52 MB
16	31.98 μs CV 0.36% · H 0.52 MB	25.00 μs CV 0.19% · H 1.05 MB	26.16 μs CV 0.43% · H 2.10 MB
32	25.51 μs CV 0.19% · H 2.10 MB	26.24 μs CV 0.41% · H 4.19 MB	32.10 μs CV 1.16% · H 8.39 MB
64	26.10 μs CV 0.48% · H 8.39 MB	33.41 μs CV 0.95% · H 16.78 MB	45.25 μs CV 0.72% · H 33.55 MB
96	31.28 μs CV 1.00% · H 18.87 MB	50.07 μs CV 0.16% · H 37.75 MB	50.64 μs CV 0.10% · H 75.50 MB
128	41.43 μs CV 0.72% · H 33.55 MB	50.48 μs CV 0.11% · H 67.11 MB	48.23 μs CV 1.10% · H 134.22 MB
192	50.60 μs CV 0.09% · H 75.50 MB	48.69 μs CV 0.66% · H 150.99 MB	138.28 μs CV 0.35% · H 301.99 MB
256	51.36 μs CV 1.46% · H 134.22 MB	64.77 μs CV 0.21% · H 268.44 MB	212.43 μs CV 0.08% · H 536.87 MB

FP32 host-resident latency

np	nh=512	nh=1024	nh=2048
8	30.24 μs CV 0.20% · H 0.13 MB	31.97 μs CV 0.28% · H 0.26 MB	34.76 μs CV 0.19% · H 0.52 MB
16	32.82 μs CV 0.16% · H 0.52 MB	30.81 μs CV 0.21% · H 1.05 MB	31.45 μs CV 0.26% · H 2.10 MB
32	31.23 μs CV 0.20% · H 2.10 MB	31.63 μs CV 0.21% · H 4.19 MB	33.64 μs CV 0.18% · H 8.39 MB
64	33.62 μs CV 0.14% · H 8.39 MB	35.38 μs CV 0.17% · H 16.78 MB	37.80 μs CV 0.12% · H 33.55 MB
96	38.06 μs CV 0.21% · H 18.87 MB	41.34 μs CV 0.14% · H 37.75 MB	43.83 μs CV 0.18% · H 75.50 MB
128	43.95 μs CV 0.19% · H 33.55 MB	46.89 μs CV 0.10% · H 67.11 MB	53.40 μs CV 0.04% · H 134.22 MB
192	60.05 μs CV 0.11% · H 75.50 MB	66.59 μs CV 0.08% · H 150.99 MB	151.67 μs CV 0.05% · H 301.99 MB
256	85.31 μs CV 0.12% · H 134.22 MB	96.95 μs CV 0.32% · H 268.44 MB	243.76 μs CV 0.08% · H 536.87 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA RTX PRO 6000 Blackwell Workstation Edition (4 GPUs)

2-GPU configuration · FP64 result matrices

FP64 device-resident latency

np	nh=512	nh=1024	nh=2048
8	24.30 μs CV 0.26% · H 0.26 MB	25.47 μs CV 0.28% · H 0.52 MB	27.21 μs CV 0.21% · H 1.05 MB
16	25.13 μs CV 0.17% · H 1.05 MB	25.41 μs CV 0.55% · H 2.10 MB	31.27 μs CV 0.31% · H 4.19 MB
32	30.87 μs CV 0.58% · H 4.19 MB	40.46 μs CV 0.66% · H 8.39 MB	49.67 μs CV 0.05% · H 16.78 MB
64	45.99 μs CV 0.32% · H 16.78 MB	50.30 μs CV 0.69% · H 33.55 MB	66.19 μs CV 0.37% · H 67.11 MB
96	51.74 μs CV 0.23% · H 37.75 MB	48.58 μs CV 1.56% · H 75.50 MB	62.06 μs CV 0.04% · H 150.99 MB
128	51.44 μs CV 0.09% · H 67.11 MB	61.73 μs CV 0.13% · H 134.22 MB	82.92 μs CV 0.36% · H 268.44 MB
192	61.75 μs CV 0.26% · H 150.99 MB	135.18 μs CV 0.28% · H 301.99 MB	277.03 μs CV 0.09% · H 603.98 MB
256	70.32 μs CV 0.43% · H 268.44 MB	209.87 μs CV 0.08% · H 536.87 MB	394.53 μs CV 0.02% · H 1,073.74 MB

FP64 host-resident latency

np	nh=512	nh=1024	nh=2048
8	29.24 μs CV 0.16% · H 0.26 MB	29.81 μs CV 0.22% · H 0.52 MB	30.51 μs CV 0.20% · H 1.05 MB
16	30.10 μs CV 0.28% · H 1.05 MB	30.70 μs CV 0.32% · H 2.10 MB	31.43 μs CV 0.17% · H 4.19 MB
32	32.24 μs CV 0.18% · H 4.19 MB	33.88 μs CV 0.12% · H 8.39 MB	38.66 μs CV 0.17% · H 16.78 MB
64	37.36 μs CV 0.14% · H 16.78 MB	46.88 μs CV 0.14% · H 33.55 MB	61.14 μs CV 0.13% · H 67.11 MB
96	44.94 μs CV 0.24% · H 37.75 MB	46.56 μs CV 0.10% · H 75.50 MB	55.42 μs CV 0.18% · H 150.99 MB
128	49.03 μs CV 0.16% · H 67.11 MB	57.48 μs CV 0.07% · H 134.22 MB	88.93 μs CV 0.05% · H 268.44 MB
192	70.91 μs CV 0.06% · H 150.99 MB	149.25 μs CV 0.07% · H 301.99 MB	294.17 μs CV 0.04% · H 603.98 MB
256	102.69 μs CV 0.05% · H 268.44 MB	243.91 μs CV 0.03% · H 536.87 MB	432.29 μs CV 0.02% · H 1,073.74 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA RTX PRO 6000 Blackwell Workstation Edition (4 GPUs)

3-GPU configuration · FP32 result matrices

FP32 device-resident latency

np	nh=512	nh=1024	nh=2048
8	46.88 μs CV 0.87% · H 0.13 MB	50.98 μs CV 0.39% · H 0.26 MB	51.84 μs CV 0.87% · H 0.52 MB
16	51.45 μs CV 0.79% · H 0.52 MB	45.50 μs CV 1.73% · H 1.05 MB	53.21 μs CV 0.68% · H 2.10 MB
32	47.81 μs CV 0.78% · H 2.10 MB	54.06 μs CV 0.62% · H 4.19 MB	54.25 μs CV 0.34% · H 8.39 MB
64	54.37 μs CV 0.75% · H 8.39 MB	54.68 μs CV 1.29% · H 16.78 MB	54.82 μs CV 0.33% · H 33.55 MB
96	53.78 μs CV 1.31% · H 18.87 MB	53.78 μs CV 1.28% · H 37.75 MB	54.51 μs CV 0.27% · H 75.50 MB
128	54.80 μs CV 1.58% · H 33.55 MB	54.48 μs CV 1.25% · H 67.11 MB	54.59 μs CV 0.39% · H 134.22 MB
192	54.66 μs CV 0.87% · H 75.50 MB	54.80 μs CV 6.19% · H 150.99 MB	69.00 μs CV 1.07% · H 301.99 MB
256	55.48 μs CV 1.09% · H 134.22 MB	68.93 μs CV 1.46% · H 268.44 MB	164.47 μs CV 0.22% · H 536.87 MB

FP32 host-resident latency

np	nh=512	nh=1024	nh=2048
8	37.69 μs CV 0.51% · H 0.13 MB	39.13 μs CV 0.28% · H 0.26 MB	41.93 μs CV 3.29% · H 0.52 MB
16	39.44 μs CV 0.65% · H 0.52 MB	38.70 μs CV 3.56% · H 1.05 MB	39.65 μs CV 0.12% · H 2.10 MB
32	39.34 μs CV 0.20% · H 2.10 MB	40.14 μs CV 0.18% · H 4.19 MB	41.32 μs CV 0.16% · H 8.39 MB
64	42.58 μs CV 0.14% · H 8.39 MB	44.60 μs CV 0.16% · H 16.78 MB	47.48 μs CV 0.13% · H 33.55 MB
96	46.11 μs CV 0.19% · H 18.87 MB	48.53 μs CV 0.15% · H 37.75 MB	50.57 μs CV 0.10% · H 75.50 MB
128	52.08 μs CV 0.20% · H 33.55 MB	56.01 μs CV 0.09% · H 67.11 MB	58.54 μs CV 4.36% · H 134.22 MB
192	67.31 μs CV 0.06% · H 75.50 MB	72.57 μs CV 0.05% · H 150.99 MB	78.10 μs CV 0.13% · H 301.99 MB
256	92.52 μs CV 4.10% · H 134.22 MB	99.51 μs CV 0.14% · H 268.44 MB	197.75 μs CV 0.05% · H 536.87 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA RTX PRO 6000 Blackwell Workstation Edition (4 GPUs)

3-GPU configuration · FP64 result matrices

FP64 device-resident latency

np	nh=512	nh=1024	nh=2048
8	46.61 μs CV 0.89% · H 0.26 MB	50.62 μs CV 2.51% · H 0.52 MB	52.66 μs CV 1.03% · H 1.05 MB
16	51.46 μs CV 1.40% · H 1.05 MB	52.72 μs CV 0.30% · H 2.10 MB	52.41 μs CV 0.48% · H 4.19 MB
32	52.75 μs CV 0.52% · H 4.19 MB	52.89 μs CV 0.46% · H 8.39 MB	53.02 μs CV 0.65% · H 16.78 MB
64	53.60 μs CV 0.98% · H 16.78 MB	53.31 μs CV 0.99% · H 33.55 MB	65.03 μs CV 0.58% · H 67.11 MB
96	53.40 μs CV 1.18% · H 37.75 MB	66.04 μs CV 1.03% · H 75.50 MB	92.53 μs CV 1.02% · H 150.99 MB
128	53.11 μs CV 1.70% · H 67.11 MB	67.30 μs CV 1.75% · H 134.22 MB	76.94 μs CV 4.12% · H 268.44 MB
192	68.24 μs CV 1.50% · H 150.99 MB	76.91 μs CV 0.25% · H 301.99 MB	188.77 μs CV 0.20% · H 603.98 MB
256	79.02 μs CV 1.84% · H 268.44 MB	176.67 μs CV 0.26% · H 536.87 MB	290.11 μs CV 0.22% · H 1,073.74 MB

FP64 host-resident latency

np	nh=512	nh=1024	nh=2048
8	37.58 μs CV 0.29% · H 0.26 MB	37.97 μs CV 0.22% · H 0.52 MB	38.53 μs CV 0.20% · H 1.05 MB
16	38.34 μs CV 0.21% · H 1.05 MB	38.54 μs CV 0.12% · H 2.10 MB	39.43 μs CV 8.29% · H 4.19 MB
32	39.33 μs CV 1.46% · H 4.19 MB	40.48 μs CV 0.28% · H 8.39 MB	43.06 μs CV 1.09% · H 16.78 MB
64	45.11 μs CV 0.20% · H 16.78 MB	50.15 μs CV 0.14% · H 33.55 MB	60.66 μs CV 0.13% · H 67.11 MB
96	51.91 μs CV 0.13% · H 37.75 MB	64.88 μs CV 3.86% · H 75.50 MB	86.77 μs CV 0.06% · H 150.99 MB
128	56.51 μs CV 0.19% · H 67.11 MB	62.76 μs CV 0.10% · H 134.22 MB	74.37 μs CV 0.09% · H 268.44 MB
192	75.22 μs CV 0.06% · H 150.99 MB	88.32 μs CV 0.07% · H 301.99 MB	201.33 μs CV 0.04% · H 603.98 MB
256	111.38 μs CV 0.07% · H 268.44 MB	212.97 μs CV 0.45% · H 536.87 MB	325.46 μs CV 0.02% · H 1,073.74 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA RTX PRO 6000 Blackwell Workstation Edition (4 GPUs)

4-GPU configuration · FP32 result matrices

FP32 device-resident latency

np	nh=512	nh=1024	nh=2048
8	52.96 μs CV 1.33% · H 0.13 MB	56.54 μs CV 1.46% · H 0.26 MB	56.60 μs CV 0.97% · H 0.52 MB
16	56.86 μs CV 1.21% · H 0.52 MB	56.64 μs CV 1.21% · H 1.05 MB	57.09 μs CV 0.65% · H 2.10 MB
32	51.41 μs CV 0.57% · H 2.10 MB	57.03 μs CV 0.61% · H 4.19 MB	59.50 μs CV 0.63% · H 8.39 MB
64	56.51 μs CV 0.59% · H 8.39 MB	59.98 μs CV 0.76% · H 16.78 MB	60.13 μs CV 0.58% · H 33.55 MB
96	60.40 μs CV 0.77% · H 18.87 MB	60.45 μs CV 0.55% · H 37.75 MB	60.72 μs CV 0.94% · H 75.50 MB
128	60.10 μs CV 0.90% · H 33.55 MB	60.09 μs CV 0.98% · H 67.11 MB	59.46 μs CV 0.99% · H 134.22 MB
192	59.86 μs CV 0.52% · H 75.50 MB	59.64 μs CV 0.69% · H 150.99 MB	66.85 μs CV 1.83% · H 301.99 MB
256	60.28 μs CV 0.63% · H 134.22 MB	65.97 μs CV 0.92% · H 268.44 MB	82.99 μs CV 0.38% · H 536.87 MB

FP32 host-resident latency

np	nh=512	nh=1024	nh=2048
8	44.84 μs CV 0.12% · H 0.13 MB	46.79 μs CV 0.69% · H 0.26 MB	50.24 μs CV 0.37% · H 0.52 MB
16	46.69 μs CV 0.10% · H 0.52 MB	50.35 μs CV 0.13% · H 1.05 MB	47.86 μs CV 0.08% · H 2.10 MB
32	48.14 μs CV 0.13% · H 2.10 MB	48.36 μs CV 0.15% · H 4.19 MB	49.15 μs CV 0.15% · H 8.39 MB
64	49.96 μs CV 0.13% · H 8.39 MB	51.50 μs CV 0.13% · H 16.78 MB	55.48 μs CV 0.12% · H 33.55 MB
96	54.80 μs CV 0.18% · H 18.87 MB	57.77 μs CV 0.10% · H 37.75 MB	58.79 μs CV 0.08% · H 75.50 MB
128	60.15 μs CV 0.13% · H 33.55 MB	62.58 μs CV 0.11% · H 67.11 MB	64.74 μs CV 0.07% · H 134.22 MB
192	76.77 μs CV 0.05% · H 75.50 MB	79.27 μs CV 0.11% · H 150.99 MB	85.17 μs CV 0.06% · H 301.99 MB
256	99.61 μs CV 0.09% · H 134.22 MB	105.77 μs CV 0.05% · H 268.44 MB	122.11 μs CV 0.09% · H 536.87 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

RunPod · NVIDIA RTX PRO 6000 Blackwell Workstation Edition (4 GPUs)

4-GPU configuration · FP64 result matrices

FP64 device-resident latency

np	nh=512	nh=1024	nh=2048
8	51.24 μs CV 0.41% · H 0.26 MB	54.02 μs CV 0.41% · H 0.52 MB	58.39 μs CV 0.99% · H 1.05 MB
16	54.40 μs CV 0.61% · H 1.05 MB	57.87 μs CV 0.69% · H 2.10 MB	59.09 μs CV 0.82% · H 4.19 MB
32	58.78 μs CV 0.86% · H 4.19 MB	59.12 μs CV 0.78% · H 8.39 MB	58.98 μs CV 0.68% · H 16.78 MB
64	59.25 μs CV 0.52% · H 16.78 MB	59.40 μs CV 2.27% · H 33.55 MB	69.96 μs CV 0.79% · H 67.11 MB
96	59.23 μs CV 0.82% · H 37.75 MB	73.12 μs CV 0.40% · H 75.50 MB	92.17 μs CV 0.51% · H 150.99 MB
128	68.87 μs CV 1.32% · H 67.11 MB	81.52 μs CV 0.29% · H 134.22 MB	110.52 μs CV 0.31% · H 268.44 MB
192	64.70 μs CV 0.95% · H 150.99 MB	70.72 μs CV 0.93% · H 301.99 MB	155.39 μs CV 0.13% · H 603.98 MB
256	73.48 μs CV 0.46% · H 268.44 MB	104.63 μs CV 0.39% · H 536.87 MB	231.63 μs CV 0.29% · H 1,073.74 MB

FP64 host-resident latency

np	nh=512	nh=1024	nh=2048
8	46.34 μs CV 0.09% · H 0.26 MB	46.66 μs CV 0.22% · H 0.52 MB	47.14 μs CV 0.15% · H 1.05 MB
16	46.55 μs CV 0.15% · H 1.05 MB	46.99 μs CV 0.07% · H 2.10 MB	48.23 μs CV 0.09% · H 4.19 MB
32	47.64 μs CV 0.10% · H 4.19 MB	48.43 μs CV 0.18% · H 8.39 MB	50.30 μs CV 0.18% · H 16.78 MB
64	52.19 μs CV 0.16% · H 16.78 MB	56.55 μs CV 0.09% · H 33.55 MB	64.69 μs CV 0.06% · H 67.11 MB
96	58.34 μs CV 0.06% · H 37.75 MB	68.64 μs CV 0.08% · H 75.50 MB	85.60 μs CV 0.05% · H 150.99 MB
128	71.07 μs CV 0.07% · H 67.11 MB	85.38 μs CV 0.04% · H 134.22 MB	114.94 μs CV 0.08% · H 268.44 MB
192	82.45 μs CV 0.06% · H 150.99 MB	91.41 μs CV 0.06% · H 301.99 MB	172.47 μs CV 0.05% · H 603.98 MB
256	110.01 μs CV 0.04% · H 268.44 MB	142.23 μs CV 0.03% · H 536.87 MB	272.87 μs CV 0.15% · H 1,073.74 MB

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

AWS f2.6xlarge · AMD/Xilinx Virtex UltraScale+ VU47P (1 FPGA)

Hardware		Software and measurement	
Accelerator	AMD/Xilinx Virtex UltraScale+ VU47P	Execution environment	AWS f2.6xlarge
Compute resources	1 FPGA	Operating system	Ubuntu 24.04.1 LTS
Shell	AWS F2 Small Shell	Kernel	6.8.0-1021-aws
Shell version	0x10212415	AWS FPGA SDK	2.3.2
Host processor	AMD EPYC 7R13 Processor	Test coverage	3 fixed-bitstream cases · FP32 · np 16-32 · nh 1024-2048
Host architecture	x86-64	Measurement	5 warm-up invocations · 30 measured invocations · 8 steps per invocation
		Timing boundary	Host-observed COMPUTE_COMMIT_STEP
		H residency	Loaded before timing
		Numerical accuracy	3/3 passed · max abs 5.69e-03 · max rel 2.48e-07

Reference workload. At np=32, nh=1024, the fixed-bitstream prepared-step boundary is 12.49 μ s in FP32.

Measured fixed-bitstream performance

These FP32 measurements use the FPGA-specific compute-and-commit timing boundary. Each tested configuration passes its recorded numerical check.

3/3 Test cases passed	FP32 Fixed-bitstream precision	1.59× np32 / np16 latency
Fixed-bitstream workload	Median (μ s)	CV
np= 16, nh=1024	7.87	1.28%
np= 16, nh=2048	7.86	0.64%
np=32, nh=1024	12.49	1.34%

Latency remains nearly unchanged when history length doubles at np=16. Doubling nh from 1024 to 2048 changes measured latency by -0.12%, whereas doubling port count to np=32 increases latency to 1.59 times the original value. These results apply to the three tested configurations.

AWS f2.6xlarge · AMD/Xilinx Virtex UltraScale+ VU47P (1 FPGA)

FP32 fixed-bitstream result matrix

FP32 FPGA prepared-step latency

np	nh=1024	nh=2048
16	7.87 μs CV 1.28% · H 1.05 MB	7.86 μs CV 0.64% · H 2.10 MB
32	12.49 μs CV 1.34% · H 4.19 MB	— Not measured

Darker cells indicate longer latency within each table. H denotes coefficient storage in decimal MB. Timing definitions: page 4.

FPGA scope. This page reports 3 measured FP32 fixed-bitstream cases (np=16, nh=1024, np=16, nh=2048, np=32, nh=1024). FP64 is not present in this bitstream, and the FPGA boundary should not be read as a CPU/GPU timing ranking. These cases characterize only the supplied fixed bitstreams. Numerical error is evaluated against each case's reference and tolerance, not ranked against CPU/GPU errors with different precisions and checks.



Time-Domain System Equivalent

Linear dynamics, solved faster.

WEBSITE

tdse.tech

EMAIL

support@tdse.tech