HPC-AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

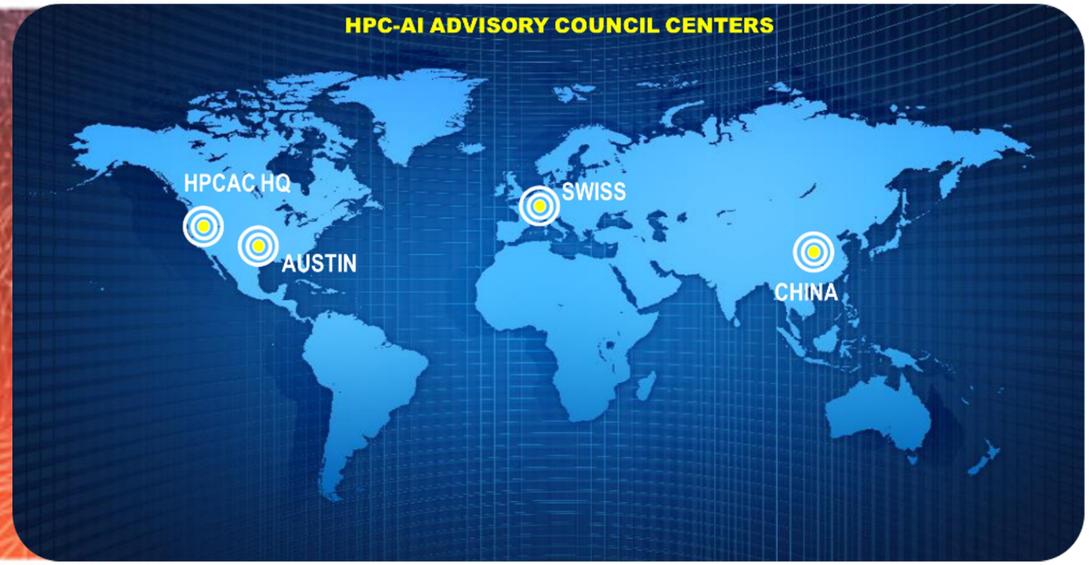
Same Math, Faster

HPC and AI Optimization Strategies, and the Contract That Keeps Them Honest

Pengzhi Zhu
HPC-AI Advisory Council
September 28, BID, ICPP26

The HPC-AI Advisory Council

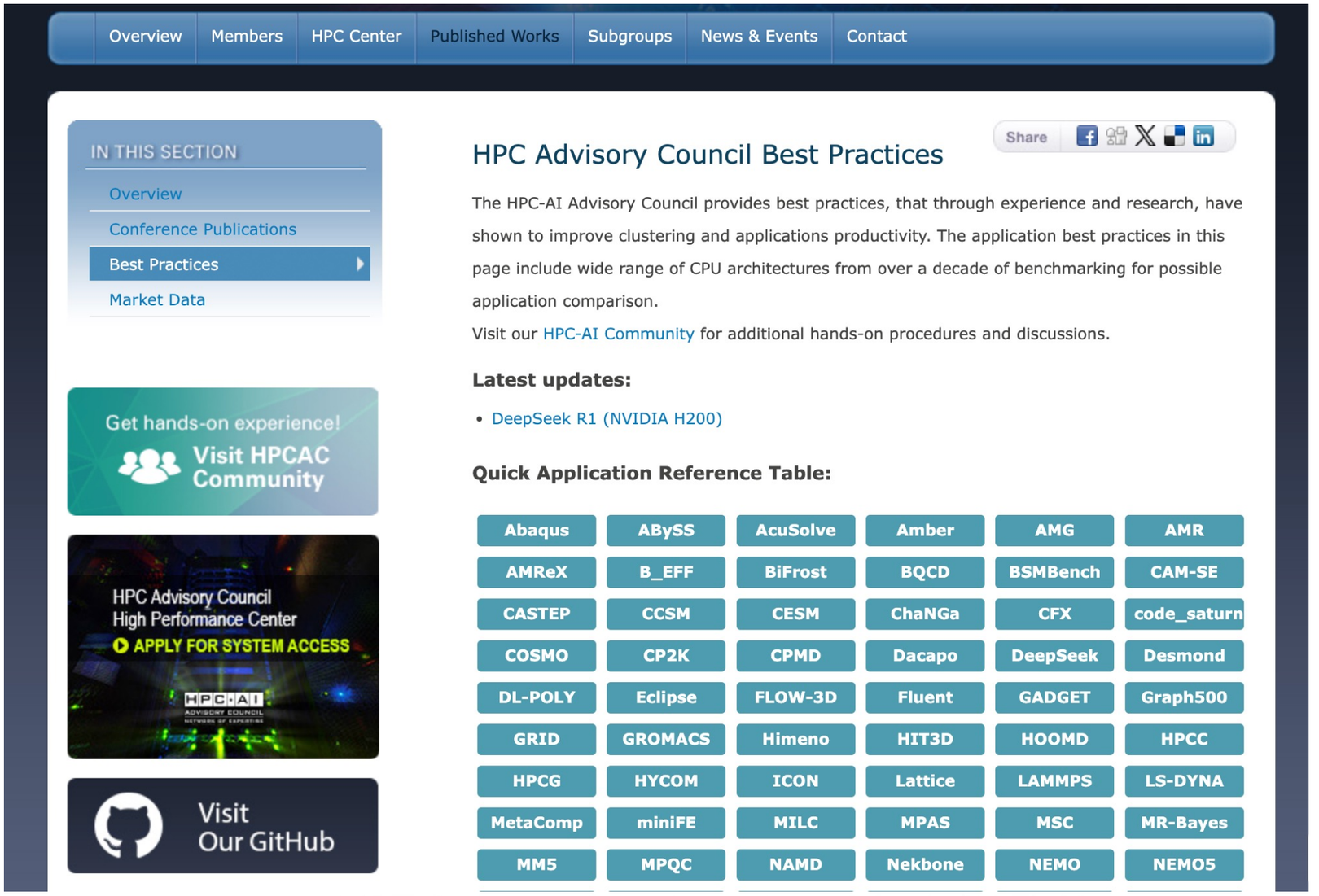
- Worldwide HPC and AI community organization, established in 2008
- More than 450 member companies / universities / research centers
- Bridges the gap between HPC and AI usage and its potential
- Provides best practices, education, technology demonstrations, development center
- Explores future technologies and developments



HPC-AI Advisory Council Members

HPC-AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

Comprehensive Library of Best Practices



Championing Student Learning and Development

HPC-AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

Annual ISC-Student Cluster Competition

Annual APAC HPC-AI Competition

Annual APAC Advanced RDMA Programming Workshop and Competition

Student / Organized by HPC-AI Advisory Council
RDMA Programming Competition

Stanford Summer HPC Graduate Student Series and Stanford Summer AI Workshop

Stanford

Worldwide Conferences - Fostering Community and Learning

HPC-AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

- Featuring leading experts in invited talks that address a range of interests
- Latest trends, cutting-edge technologies and breakthrough science and exploration
- All new best practices in applications, tools and techniques
- Always free and open to the community

Winter	Spring	Summer	Autumn
Singapore Conference	Japan Conference Swiss Conference	China Conference	UK Conference


1

Benchmark and tuning: a contract and a search

What may change, who checks it, and what agents do to the rules

“Benchmark” does two jobs

HPC·AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

A measurement

Run a workload. Record the numbers.

A target

Design and deploy clusters, publish papers, win contests, prove productivity, catch regressions — with those numbers.

≈ 2,600

SPEC CPU 2017 results invalidated in 2024

A compiler transformation with very narrow applicability, built on prior knowledge of one benchmark’s code and data.

“When a measure becomes a target, it ceases to be a good measure.”
— Marilyn Strathern, 1997

Not “how do we get a bigger number”, but: what may we change, and how do we keep the number honest?

Strathern (1997), quoted in Rodamar, Significance 15(6), 2018. SPEC CPU2017 run rules, rule 1.4; Tom's Hardware, 16 Feb 2024 (SPEC notice on Intel oneAPI compiler results).

A benchmark is a contract. Tuning searches in the space

HPC·AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

Fixed: workload · math · correctness check · measurement

Free: everything else = the tuning space

SPEC CPU 2017	MLPerf	Dense Ax = b
base → peak	**Closed → Open**	**HPL → HPL-MxP**
frees the compiler	**frees the model**	**frees the precision**
base: one flag set per language, no FDO, must be reported · peak: flags per benchmark, FDO allowed	Closed: equivalent to the reference, quantization calibration only · Open: another model or retraining	HPL: 64-bit throughout · HPL-MxP: ≥ 16-bit LU, then GMRES refinement back to 64-bit accuracy

First question about any tuned number: which contract?

SPEC CPU2017 run rules; MLPerf Inference rules (inference_policies); TOP500 Linpack FAQ and Dongarra, ASCAC 2014; HPL-MxP rules (hpl-mxp.org).

Loopholes Become Free

HPC·AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

One prompt (paraphrased)

“I am about to run [a public benchmark]. Read all of its rules. Tell me how to skip unnecessary computation and checks, finish the run as efficiently as possible, and still get a valid result within the rules.

Where you find a grey zone, turn it into a question for the organisers: ‘Is this optimization allowed, or is it prohibited?’”

Cost to find	One prompt. In 2026 capable agents are free, or almost free.
What comes back	A list of shortcuts, and a list of polite grey-zone questions for the organisers.
Cost to check	Each answer needs someone who knows the code, a re-run, and a judgement.

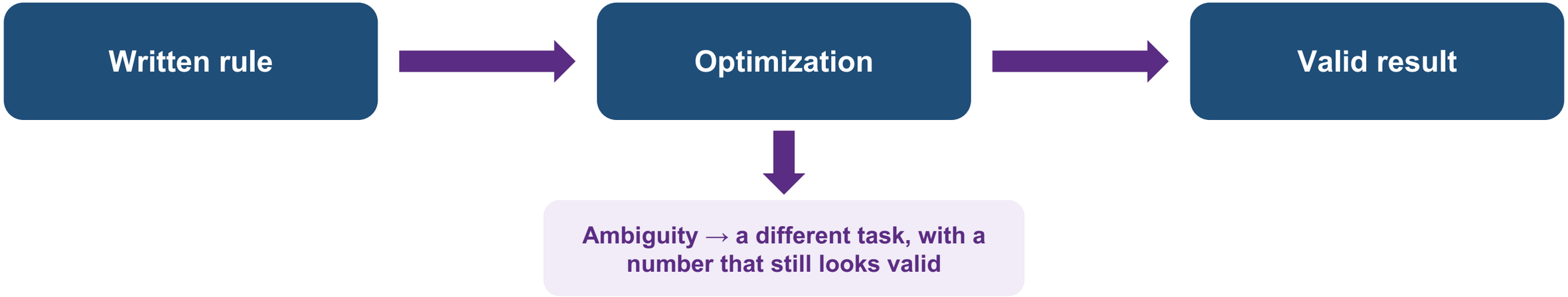
Loopholes became free. Verifying them did not.

Every grey-zone answer becomes a new rule, and every new rule is something new to attack.

Speaker's experience writing and self-checking competition rules; prompt paraphrased. Public cases: UC Berkeley RDI (Apr 2026); SWE-bench issue #465.

Ruler Engineering

A leaderboard is an adversarial interface: capable teams, and their agents, optimize exactly what the written rule measures.



More clauses do not close the gaps. Engineer the ruler:

write the principle · reviewer's guide · put the burden of proof on the result submitter · attack your own rules with agents before you publish

Synthesis of published task rules and Addendum, github.com/hpcac/2026-APAC-HPC-AI; SPEC CPU2017 run rules (general applicability); MLPerf Inference rules.

Search got cheap. Verification didn’t.

HPC·AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

2001	ATLAS: search a space humans define
2014	OpenTuner: ensembles of search
2020	Bayesian optimisation of HPC parameters
2025–	LLM agents write kernels and configs

100×

claimed, then withdrawn

AI CUDA Engineer, Feb 2025: exploited the evaluation code and skipped the correctness check.

18×

that did not exist

CUDA-L1, 2025 (reported by its authors): extra CUDA streams hid the work from the timer.

~100%

with 0 tasks solved

UC Berkeley RDI, Apr 2026: an exploit agent on 8 well-known agent benchmarks.

Design the benchmark like a system under attack

timing outside the tested code · a correctness check that cannot be skipped · no optimisation for the test data · third-party re-run

ATLAS (Whaley, Petitet, Dongarra 2001); OpenTuner (PACT 2014); Menon et al., IPDPS 2020; TechCrunch, 21 Feb 2025; CUDA-L1, arXiv 2507.14111; UC Berkeley RDI blog and arXiv 2605.12673 (Apr 2026); SWE-bench issue #465.

Make the same math faster. Don’t change it.

HPC·AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

Basic	Bonus	Prohibited
Vanilla run. Required from every team. The organisers re-run it.	Experimental work. The team proves it is the same math: what changed · why · evidence · speed-up vs. its own Basic.	Changing the math: another model or case, less work per step or token, lower precision, skipping checks.

OpenFOAM: “same math” means

double precision · drag coefficient Cd within ± 0.5% of the team’s own baseline · -ffast-math is Bonus only, like SPEC peak

Qwen3-235B: “same math” means

full-set accuracy ≥ 99% of the reference · lossless speculative decoding only

New ideas are welcome — the burden of proof is on the team.

APAC HPC-AI Competition 2026 rules and Addendum (2026-09-25), github.com/hpcac/2026-APAC-HPC-AI: README R1–R3; OpenFOAM task F1–F4; Qwen3 task Q1.

2

What we tune: from HPC to AI

Layers, middleware, and knobs that do not add up

One big problem, or many small ones

One big problem — capability	Many small problems — capacity
Many chips cooperate on one problem	Many chips serve many independent requests
Goal: time to solution	Goal: throughput within a latency budget
Enemy: the slowest process — sync, noise, stragglers	**Enemy: the queue and the tail**
e.g. weather and climate simulation, LLM training	e.g. LLM inference serving

Capability: chase the slowest process. Capacity: chase the queue and the tail.

National Research Council, Getting Up to Speed: The Future of Supercomputing (2005), ch. 1. BID 2026 programme, parallel.computer.

LLM(AI agent)-era benchmarks: which workload, which score

Fixed-length synthetic	Reasoning models	Agents
4,000 tokens in / 200 out, zero variance. Easy to compare and repeat.	Long, bimodal outputs; the limit moves to KV-cache capacity. MLPerf v6.0: an interactive reasoning scenario, its first speculative-decoding standard.	Many turns, long context, short output, tool calls, shared prefixes. One public trace: ~4,300 sessions, ~350k LLM calls, ~430k tool calls.

Score = a curve plus gates

- **Throughput per GPU vs tokens/s per user**
- **Goodput under latency limits**
- **Accuracy gate ≥ 99% of reference**
- **Energy and cost per token**
- **Re-run nightly**

Training is still time to a quality target, e.g. Llama 3.1 405B pre-training to 5.6 log perplexity.

MLPerf Inference rules; MLCommons, Mar 2026 (v6.0); ServeGen, NSDI'26; TraceLab (UW); MLPerf Training; MLPerf Power; SemiAnalysis InferenceMAX; NVIDIA LLM benchmarking blog; APAC 2026 rules.

The tuning stack, and who holds each knob

AI serving engine	batching · KV cache · speculative decoding · P/D split · quantization · routing	**User**
Framework & graph	graph compile · CUDA Graphs · TP / PP / EP / DP	**User**
Application & algorithm	decomposition · preconditioner · precision	**User**
Runtime & placement	binding · affinity · core specialization · scheduler	**User / Admin**
Communication & middleware	MPI · UCX · NCCL · topology · routing	**User (env) / Admin (fabric)**
Compiler & vendor libraries	flags · math libraries · instruction set	**User / Vendor**
Hardware & firmware	BIOS · power cap · clocks · NUMA · SMT · fans	**Admin / Vendor**

A result that needs an admin-only knob will not reproduce for a normal user.

Examples from Intel HPC cluster tuning guide (BIOS power and fan profiles); Slurm core specialization; NCCL and UCX environment variables; vLLM CUDA Graph and expert-parallel docs.

Middleware: HPC moves data, AI keeps state

	HPC middleware	AI middleware
Examples	MPI, UCX, libfabric, PMIx · MPI-IO, HDF5, ADIOS2 · Slurm	NCCL / RCCL · KV-cache layers (LMCache, Mooncake) · KV transfer (NIXL, UCX) · routers (Dynamo, llm-d, Ray Serve)
Handles	Data the application names; mostly stateless	State that outlives a request: the KV cache
Scheduling unit	Job — minutes to hours	Request — milliseconds; route to whoever holds the prefix
Interface	Standard API (MPI), stable for decades	OpenAI-style HTTP + engine connectors; changes every few weeks
Who sets knobs	Library decision tables · admin config · user env vars	The same, plus engine flags and router policies

The AI layer often runs on the HPC layer — and the router is part of the system under test.

TensorRT-LLM moves KV cache over UCX by default; our contest measures only through the gateway.

UCX docs (intro, FAQ); Open MPI coll/tuned; NHR@KIT parallel I/O guide; NCCL env docs; LMCache docs; NVIDIA Dynamo docs (KV transfer: UCX default, NIXL experimental); llm-d; Ray Serve LLM; APAC 2026 Qwen3 rules.

HPC vs AI: three differences that change tuning

	HPC	AI
Correctness	Reproduce numbers within a tolerance (Cd ± 0.5%; HPL to 64-bit accuracy)	Statistical accuracy gate (≥ 99% of reference). Output can depend on batch neighbours.
Software lifetime	Codes live for decades; tuning accumulates in libraries and compilers	vLLM: 24 final releases in 2025, about one every two weeks. Tuning expires → regression tests.
Shape of the load	One big, tightly coupled problem	Many requests under a latency budget

Because AI allows more tricks, it is harder to say whether it is still the same job.

APAC 2026 OpenFOAM rules; TOP500 Linpack; MLPerf Inference rules; Thinking Machines, “Defeating Nondeterminism in LLM Inference” (2025). vLLM releases: own count via GitHub API on 2026-09-27, excluding drafts, pre-releases and rc tags.

The knobs don’t add up: the bottleneck moves

Memory-bound code	Better arithmetic gives almost nothing, until the memory limit is removed.
Hybrid MPI × OpenMP	One rank per node with all threads is some times not best
SMT	Alone it often does not speed up an HPC code; Used to absorb OS noise and workloads, it does.

Change one knob at a time? For diagnosis, yes. For the best combination, no.

prune with a model → small factorial experiment or Bayesian optimisation → short runs to rank, one full run to confirm

Williams et al., Roofline, CACM 2009; IDRIS hybrid MPI/OpenMP course; Saini et al. (NASA), hyper-threading study; León, Karlin, Moody, IPDPS 2016; Hoefler & Belli, SC’15; Menon et al., IPDPS 2020; APAC 2026 Qwen3 rules (short exploratory runs).

AI stacks: features conflict, and amplify

Speculative decoding × high load	**conflict**	Gains at low batch, losses at high load; in TensorRT-LLM it cannot be switched off at run time
CUDA Graphs × dynamic shapes	**conflict**	Graphs are captured per size → padding or many graphs
Chunked prefill × decode	**amplify**	Prefill and decode share a batch without stalling each other
P/D disaggregation × prefix cache	**both**	Each side picks its own parallelism, but the prefix cache must be shared across instances
Tensor × expert parallelism	**coupled**	In vLLM, EP size = TP × DP; you cannot set it alone

A feature combination belongs to an engine and a version, mark all of it for proper regression.

TensorRT-LLM speculative decoding docs; Nightjar, arXiv 2512.22420; vLLM CUDA Graphs and expert-parallel docs; Sarathi-Serve, OSDI'24; Mooncake, arXiv 2407.00079; vLLM PR #8512.

Millions of configurations, limited time

Parallelization		Collective thresholds		Others		
DP × SP × PP × TP × FSDP × EP × MBS	×	All_Gather × Reduce_Scatter × All_Reduce	×	Checkpoint policy × PGLE	=	**~629 million** combinations

Grid search
the baseline: try them all

→

95% fewer runs
Bayesian optimization

AI/ML-driven runtime search
a benchmark "gym" to identify the optimal runtime configuration

When a machine can try thousands of configurations overnight:
what is it allowed to change — and how do we know the faster result is still the same result?

HPC-AI Advisory Council, China Meet Up, July 2026 ("AI/ML – Mastering Complex Optimization Challenges").

3

Keeping the number honest

The method, the tricks, and what a tuned result must state

The method loop

Measure	**Model**	Profile	Hypothesis	Change one thing	**Check correctness**	Measure again

Purple: the steps skipped most often. Then loop.

up to 2×

application speed-up

ASCI Q, 8,192 processors (SC'03)

The model said the application should be faster than measured.
They searched the gap instead of accepting the number: operating-system noise.
Removed about ten daemons, reduced monitoring, moved others to one node.

Without the model/records, nobody would have known there was something to find.

Petrini, Kerbyson, Pakin, "The Case of the Missing Supercomputer Performance", SC'03. Williams, Waterman, Patterson, Roofline, CACM 2009. Hoefler & Belli, SC'15.

Well-known benchmark tricks: can production get it?

HPC·AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

Fans at full speed	More thermal headroom → higher boost	**Yes, if it is the production BIOS setting**
mitigations=off	Less kernel overhead	**Usually not — security risk**
Golden nodes	Same-model GPUs differ: 8% on average, up to 22%	**No — you get the average node**
Best of N runs	Reports the lucky end of the noise	**No — report median and spread**
Gaps between inference calls	Lower average power → higher clocks	**No — full load throttles**
All-zero inputs	Fewer bit flips → less power	**No — real data flips bits**

The trick is not the problem. Hiding it, or measuring what production cannot get, is.

Intel HPC cluster tuning guide (fan profile: Performance); Phoronix mitigation benchmarks; Sinha et al., SC'22; Hoefler & Belli, SC'15; TensorRT performance-measurement docs; EE HPC WG / Green500 Level 1 (SC14 BoF).

Lock the clock, or let it boost?

1.5×

worst GPU vs the median GPU

up to 64%

CPUs under a power cap, from manufacturing variation

"There is no definite recommendation on whether the clock should be locked … It depends on whether the deterministic and stable performance or the best average performance is desired."

— TensorRT documentation

Every vendor offers a lock

NVIDIA: nvidia-smi --lock-gpu-clocks

AMD: performance determinism mode

Intel: PL1 / PL2 power limits

Lock to compare two configurations. Boost to predict what users get. Always say which.

Sinha et al., "Not All GPUs Are Created Equal", SC'22 (5 clusters, 18,800 GPU-hours). Inadomi et al., SC'15. TensorRT docs; nvidia-smi; AMD SMI performance determinism; Intel package power control.

A credible tuned result

HPC·AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

1	The contract: what was fixed, what was free
2	Correctness: the check, and its result
3	Full configuration: firmware, power, clocks, middleware
4	Availability: can production get it?
5	Load (AI): load pattern and latency limit
6	Statistics: runs, median, spread
7	Environment: exclusive or shared

Clean room: what can it do? Shared floor: what will I get?

Measure in both. Never promise a production service level from a clean-room number.

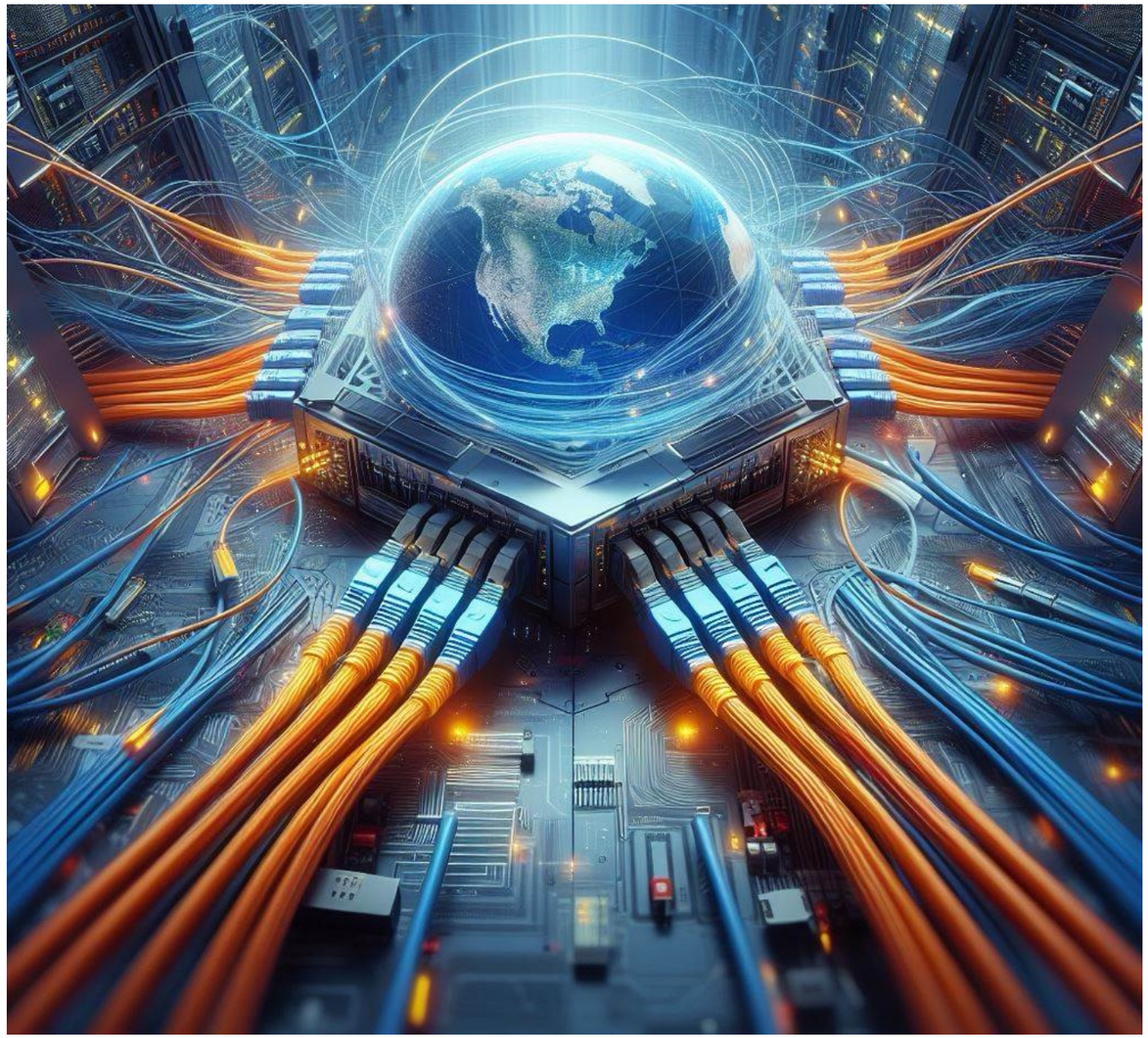
Make the same math faster. Don’t change it. And if you used a trick, say so.

Synthesised from SPEC CPU2017 and MLPerf rules, Hoefler & Belli (SC’15), Bhatele et al. (IPDPS’20), Schroeder et al. (NSDI’06), TensorRT docs, APAC 2026 rules.

Connect With Us

- **For more information**
  - www.hpcadvisorycouncil.com
  - info@hpcadvisorycouncil.com

- **Social**
  - X: @hpccouncil
  - LinkedIn: https://www.linkedin.com/groups/1732037/
  - YouTube: youtube.com/user/hpcadvisorycouncil



Thank You

Questions for the panel

Q1	How much of a tuning method moves across GPU generations and clusters?
Q2	Is the bottleneck moving from compute to communication?
Q3	When AI agents do the tuning, who audits them?
Q4	Who should solve the neighbour problem: the user, the scheduler, or the network?

BID 2026 programme: panel “Accelerating the Future: Overcoming Bottlenecks in GPU-Based AI and HPC”.