COMPUTING INFRASTRUCTURE BREAKOUT

Where does HEP computing go from here?

CLARIPHY AI Collaboration Meeting • Madison, WI • Sept 14–16, 2026

Session Organizers

Daniel Diaz (SDSC / NRP / CMS) • Ian Fisk (Flatiron Institute) • Rob Gardner (UChicago / ATLAS) • Oliver Gutsche (FERMILAB)

Session format (90 minutes)

0:00

15 min

Framing talk

Where the current model came from, why it's under strain, and the questions on the table.

0:15

55 min

Guided discussion

Open, facilitated discussion through the focus questions — live-polled with Slido.

1:10

15 min

Synthesize

Organizers and the room pull out the top 3–4 themes together, live on screen.

1:25

5 min

Wrap

Today’s computing model: distributed, pledged, tiered

The WLCG tier-1 / tier-2 / tier-3 structure is the resource mode. Institutions pledge dedicated hardware, sized for steady-state simulation, reconstruction, and analysis at LHC data rates.

- Built around dedicated, owned hardware pledges per institution
- Sized for predictable, **batch-style** CPU-intensive work
- Worked well when compute and storage costs were falling year over year

That model is now colliding with hardware prices

3–5×

rise in DRAM prices since early 2025: the AI memory shortage, expected to persist to 2030

35%

of system bill of materials is now memory, up from 15–18% a quarter earlier

~2 yr

backorder on enterprise disk; 2026 supply was booked out before February

The tier model assumed ~170 institutions could each keep buying the same hardware every year, and that falling $/unit would turn flat money into growing capacity. Both halves of that assumption broke at once.

And the scale gap is widening, not closing

2026 AI capital spend: four hyperscalers vs. DOE's entire Genesis call

$760B

Microsoft, Alphabet, Meta, Amazon combined capex

$293M

DOE Genesis Mission FY26 funding call · a ratio of ~2,600 : 1

20–40×

More CPU and disk that ATLAS and CMS need at HL-LHC under current computing models. Flat budgets deliver roughly 2× over the same period — a shortfall of 10–20×.

HL-LHC starts June 2030. We are in Long Shutdown 3 right now — this is the window where the computing model gets decided.

Data point: What it takes to serve one frontier open-weight model

Worked example: Kimi K2.6 — 1.04T parameters, mixture-of-experts, ~32B active per token. Only 32B activate, but all 1.04T must sit in memory. GLM, Qwen and DeepSeek's large open-weight models size similarly.

8-bit (production)	**~1,123 GB**	8 × H200 (141 GB) = 1,128 GB	*the clean fit — one node*
4-bit	**~634 GB**	8 × A100/H100 80 GB = 640 GB	*minimal headroom*
FP16 / BF16	**~2,243 GB**	multi-node cluster	*not a single-box option*

One 8-way H200 node ≈ $370,000

Range $320–420K · ~8 kW · sustained 10K+ tokens/sec. Individual H200 cards run $32–40K. An 8-GPU B300 system is $550–650K with 288 GB per GPU.

The demand is changing shape, not just size

The tier model was built for one workload: CPU batch, highly parallel, latency-tolerant. Three different shapes now sit on top of it, and each wants a different machine and a different operating model.

CPU batch

Simulation, reconstruction, analysis trains

Queue for hours. Scales out. Latency does not matter.

Pledged cores

GPU training

Foundation models, surrogate simulation

Dense multi-GPU nodes, scheduled, memory-bound.

Hard to pledge

GPU inference

Reconstruction, triggering, tagging

Bursty, latency-bound, wants to be a service.

Service-shaped

Interactive / agentic

LLM access, agentic workflows, coding

Always on, sub-second, metered per token.

Budget-shaped

Both agencies are consolidating — at the same time

DOE[†]

- HEP research line: $431M → $283M in the FY27 request (−34%)
- Genesis Mission: 20 HEP awards from 381 applications (5.2%)
- Awards are large, multi-institution, industry-partnered
- No new major construction starts expected for 6–7 years

NSF[‡]

- ~6,100 new grants in FY26 — lowest since the early 1980s
- 46% below the 2021–2024 average
- Cutting the number of solicitations by half or more
- Explicitly “focusing on longer awards with larger funding amounts”

The one P5 computing recommendation DOE is funding (4f: shared software, sustained computing R&D) is being delivered “in support of the DOE-wide Genesis Mission.” Computing is the surviving line.

[†]HENP for DPF July 2026

[‡]https://www.nature.com/articles/d41586-026-02574-6

A shift is already underway

Nobody in this room decided to move from pledged hardware to metered capacity. Funding agencies and industry are moving that way independently, and the community inherits the consequences.

What funding agencies are doing	DOE routes computing through Genesis and the American Science Cloud. NSF stands up NAIRR as allocations and cloud credits, not site pledges. Neither is shaped like an MoU.
What industry is doing	Capacity is sold as tokens and GPU-hours, priced per unit and metered. That is now the default commercial shape, and the one our partners build for.
What that implies for us	Allocation windows rather than owned racks. Per-experiment and per-person budgets. Explicit trade-offs about which workflows are affordable — and a real question about equitable access.

Open question for the room: is a metered model something to resist, or to negotiate good terms on?

Discussion Topic: Industry partnerships

Both agencies have made partnership a condition of the money, not a bonus. The community has decades of precedent for doing this well — the question is whether we scale it.

It is written into the awards

Every HEP Genesis selection is multi-institution. AMD, Google, NVIDIA, AWS, SuperMicro and GlobalFoundries appear across the list — as named partners, not vendors.

It is how NAIRR exists at all

~$100M of in-kind contribution from 28+ companies: NVIDIA $30M, Microsoft $20M in Azure credits, plus OpenAI, Google, AWS, Meta, IBM, Intel and Anthropic.

We already have a working model

CERN openlab has run public–private R&D since 1983 — currently Oracle, Siemens, NVIDIA, Intel, Pasqal and others, with a defined governance and cost-sharing structure.

Discussion Topic: Inference-as-a-service (e.g., SONIC)

The problem it solves	Experiment workflows are CPU-bound and latency-tolerant; the ML inside them wants a GPU. Rather than putting a GPU in every worker node, the job calls out to a shared inference server.
How it works	SONIC is the client pattern inside experiment software; SuperSONIC is the Kubernetes-based server infrastructure — Triton, load balancing, autoscaling — so the CPU-to-GPU ratio can be tuned independently of the job mix.
Who uses it	CMS for ParticleNet, DeepMET, DeepTau and ParT; ATLAS for Exa.TrkX and Traccc; IceCube for CNN event classifiers. Deployed at Purdue (Geddes, Anvil), NRP Nautilus, and the UChicago ATLAS Analysis Facility.
Why it matters here	It is a template for treating expensive hardware as a shared service rather than a per-site purchase — which is exactly the shape a metered, allocation-based model would need.

Source: SuperSONIC documentation, fastmachinelearning.org.

Discussion Topic: Joint procurements

Joint hardware procurement

- Multiple institutions/experiments co-propose for shared hardware on a single award.
- Distributed across sites and made available to the wider community.
- Aggregate demand could support larger, better-optimized procurements and avoid every institution independently buying specialized hardware.
- The attraction is that it matches what both agencies say they want to fund: fewer, larger, multi-institution awards.

More Discussion Ideas

- Where are the most promising synergies across communities?
- What concrete collaborative opportunities exist near-term?
- Are there achievable demonstrator projects to build momentum?
- What resources, infrastructure, or org. support is needed?
- What workforce / training / community needs should we address?
- Inference-as-a-service, sustaining infra, industry partnerships?

Discuss locally; share your thoughts

Scan the QR code or go to slido.com to submit responses to live polls, vote on ideas, and rank priorities anonymously as we talk. Top-voted items feed directly into our closeout slide.

Open Q&A + upvoting

Surface the ideas the room cares about most, not just the loudest voice.

Live word cloud

One-word gut check on “biggest blocker” or “biggest opportunity.”

Ranking poll

Prioritize candidate demonstrator projects in the last 15 minutes.

slido.com • #CLARIPHY-COMPUTE

Q&A

- Open for any thoughts during the discussion section.
- Add questions or comments about any of the topics and upvote
- 3 polls follow after the discussion session

Thank you

Let's keep the conversation going after Madison.

DOE: what changed, in their own words

> *"To take advantage of this rapidly developing frontier AI requires collaboration and partnerships requiring a shift in the structure and support of HENP research."*
>
> Regina Rameika, Associate Director, DOE Offices of High Energy and Nuclear Physics — DPF meeting, 23 July 2026

$1.235B → $1.120B

HEP total, FY26 enacted to FY27 request

$431M → $283M

the research line specifically — a 34% cut

20 of 381

HEP Genesis awards from applications — 5.2%

$293M

total FY26 Genesis funding call across all of DOE

6–7 years

before any new major construction start is expected

Zero

single-institution awards in the HEP Genesis selection list

Awards are large, multi-institution and industry-partnered: AMD, Google, NVIDIA, AWS, SuperMicro, GlobalFoundries all appear in the selection list. AIDA-Scout (topic 14-A) is one of them.

NSF is consolidating on the same timeline

~6,100	new grants expected in FY2026 — the lowest count since the early 1980s
−46%	against the 2021–2024 average; roughly 30% below last year alone
Half or fewer	solicitations, consolidated into broader calls routed internally
$1B	of an $8.8B budget held centrally and unavailable to core grant programs

NSF's own stated direction: “focusing on longer awards with larger funding amounts.” Both agencies moving the same way at once is why a consortium is the only fundable shape.

It is not only LLMs — the experiments are growing too

Detector ML demand is rising independently of the AI-tooling conversation, driven by more data and rarer signals.

40 MHz

L1 scouting reads out at the full collision rate — the extreme end of the throughput problem

50 ns

latency budget for a trigger decision, forcing FPGA-resident neural networks

25×

more data from HL-LHC than the LHC has produced in its entire history

3 experiments

CMS, ATLAS and IceCube already run inference-as-a-service through SuperSONIC

Model-independent searches make this worse, not better: if you cannot specify the signal in advance you cannot cut the data down in advance, so anomaly detection has to run on far more of it. That is the AIDA-Scout problem, and it is a compute and storage problem before it is a physics problem.

Sources: arXiv:2411.19506 (CMS L1 anomaly detection); SuperSONIC documentation; CERN Courier.