Perspectives on Foundation Models and Agentic Architectures for Computational Fluid Dynamics

Neil Ashton - Distinguished Engineer, NVIDIA

April 9, 2026

Advanced Modeling & Simulation (AMS) Seminar Series
NASA Ames Research Center, April 9th, 2026

Summary

- **Current State:** CFD solver (we'll discuss the rest of the workflow later) is undergoing a disruptive transformation from classical methods to AI surrogates.
- **The Advantage:** AI surrogate models predict results in under a second, orders of magnitude faster than conventional solvers. Also opens the door to truly automated shape optimization.
- **The Gap:** Most models utilize small, use-case specific datasets; generalizability remains a question. Could ChatGPT for fluids be created?
- **Hardware:** Simulation/training/inference - what is the relative importance is up for debate over the next 5-10 years.

Market Trends (2026)

- Historically, progress was driven by increasing HPC resources for higher fidelity (LES, DNS).
- GPU-accelerated solvers provided an initial order-of-magnitude reduction in runtime.
- **The AI Paradigm:** Shift from "HPC-only" to AI surrogate emulators - but how much, how soon?
- **Market Drivers:** Real-time predictions enable massive design-space exploration at low cost. But you need training data in the first place. GPU solvers are still very important.

The Industry-Scale CFD Pipeline

The process transforms high-dimensional problems into low-dimensional engineering insights via three stages:

1. **Inputs:** Translation of physics into a discrete mathematical system.
2. **Compute Engine:** Linearizing and solving large sparse linear systems ($Ax = b$).
3. **Outputs:** Distilling terabyte-scale raw data into scalar quantities (e.g., C_L, C_D).

Based on analysis from Fluid Intelligence: A Forward Look on AI Foundation Models in CFD
`https://arxiv.org/abs/2511.20455v1`

The Foundation of Numerical Methods: Meshing

- **Finite Volume Method (FVM):** Dominant in commercial codes; uses unstructured cells for geometric flexibility.
- **Refinement:** Meshes are refined in regions of high gradients (boundary layers, shock waves).
- **Scale:** Modern simulations utilize hundreds of millions to billions of cells.

The Physics Modeling Spectrum

- **RANS (Reynolds-Averaged Navier-Stokes):** Models the effects of unresolved turbulence scales; computationally affordable but limited accuracy in massive separation.
- **LES (Large Eddy Simulation):** Resolves large eddies and models sub-grid scales (SGS).
- **Cost:** LES cost is orders of magnitude higher than RANS.

Data Generation: RANS vs. WMLES

- **Implicit RANS Iteration:** Dominated by linear system solution (80%+ of time).
- **Explicit WMLES Time-Step:** Deterministic and cheap per step; total cost driven by millions of small time steps.
- **Metric f:** FLOPs per cell per step.
- $f_{RANS} \approx 40,000$; $f_{WMLES} \approx 2,800$.

Hardware: Roofline Analysis

- Suitability is quantified by **Arithmetic Intensity (I)**: FLOPs per byte moved from memory.
- **GPUs:** Designed for high I ("compute-bound") algorithms.
- Explicit Cartesian solvers/LBM map well to GPU Single Instruction, Multiple Threads (SIMT) models.
- GPU acceleration provides order-of-magnitude gains over CPUs. But main question is what is the optimum algorithm for the hardware and what is the hardware for the algorithm.

Complexity for Training AI Models

Model	Model size	Total compute
S_r^θ (LF)	$M_r \sim N_r^{\frac{\beta}{\alpha}}$	$\sim nf_e N_r^{1+\frac{\beta}{\alpha}} + f_r N_r^{1+\frac{(3+\kappa)\beta}{\gamma}}$
S_ℓ^θ (HF)	$M_\ell \sim N_\ell^{\frac{\beta}{\alpha}}$	$\sim nf_e N_\ell^{1+\frac{\beta}{\alpha}} + f_\ell N_\ell^{1+\frac{4\beta}{\gamma}}$
S_t^θ (HF-T)	$M_t \sim \tilde{N}_\ell^{\frac{\beta}{\alpha}}$	$\sim nf_e K \tilde{N}_\ell^{1+\frac{\beta}{\alpha}} + f_\ell \tilde{N}_\ell^{1+\frac{4\beta}{\gamma}}$

Transient high-fidelity (HF-T) models are hypothesized to be cheaper than time-averaged models for reaching the same target error.

Mixed Precision

- One of the biggest topics is mixed precision.
- We need to know how low can we go and how does this trend across generations.
- Again, what hardware/method is right?

The High-Dimensional Parameter Space

Simulations required for a Foundation Model split into four areas:

1. **Domains (N_d):** High-level areas (Aerospace, Automotive).
2. **Archetypes (N_a):** Base geometries (Sedan, Wide-body aircraft).
3. **Parameters (N_p):** Geometric/flow variables (AoA, Reynolds No.).
4. **Values per Parameter (N_{pv}):** Samples per parameter.

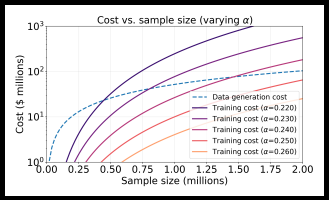
Estimated requirement: **2.25 million simulations**.

CFD Foundation Model Estimates

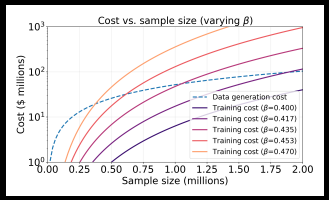
	Large	**XLarge**	**XXLarge**
Training samples (N)	2×10^5	10^6	2×10^6
Compute cost (C_c)	\$10.3M	\$51.8M	\$103.7M
Model size (M)	3.15B	56B	195B
Training cost (C^{Tc})	\$397k	\$35.5M	\$295M
Overall Field Error	–	–	**1%**

Transition from Large to XXLarge shows training costs overtaking generation costs.

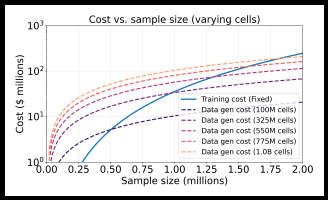
Scaling Behavior: Sensitivity Analysis



(a) Varying α



(b) Varying β



(c) Varying Cell Count

Figure 1: Comparative analysis of data generation and model training costs.

The Regime Flip

- **Small Data ($N < 10^6$):** Data generation is the primary computational burden.
- **Large Data ($N \geq 10^6$):** Coupled growth of model and dataset size causes training costs to eclipse generation.
- Immediate barrier is efficient simulation; long-term barrier is training compute.

Architecture & Inductive Biases

- **Neural Operators (FNO, GNO):** Learning mappings between infinite-dimensional function spaces.
- **Transformers:** Success in other domains (AlphaFold) now entering CFD via Attention mechanisms.
- **Physics-Informed (PINNs):** Integrate physical laws into the loss function, but not yet successful for large-scale external aerodynamics.

Market: Digital Twins & In-Situ Visuals

- High-fidelity WMLES on billion-cell grids create severe data bottlenecks.
- **In-Situ Visualization:** Analysis/rendering occurs in memory concurrently with simulation.
- Only post-processed results are saved, enabling online training of AI models.

Obstacles to a Universal Model

- **Input Space:** Geometric variation is easy to encode, but physics choices (RANS vs. LES) cannot be implicitly inferred.
- **Conditioning:** Surrogates must be explicitly conditioned on modeling choices to generalize across datasets.
- **Diversity:** Input diversity is the role determinant of performance, paralleling document diversity in LLMs.

Case Study: Dataset Composition

- **DrivAerML:** 487 simulations, Transient HRLES, 160M cells.
- **DrivAerNet++:** 8,000 simulations, Steady-state RANS, 23M cells.
- **Result:** Distinct input choices result in vastly different conditional output distributions with limited overlap.
- **Insight:** Simply combining datasets may not yield a more functional model.

Universal vs. Domain Specific Models

- Should we pursue a single Universal Model or smaller Domain-Specific ones?
- Small models trained on specific use-cases are cheaper to train and inference.
- **Open Question:** At what point does the benefit of seeing a vastly different input space saturate?

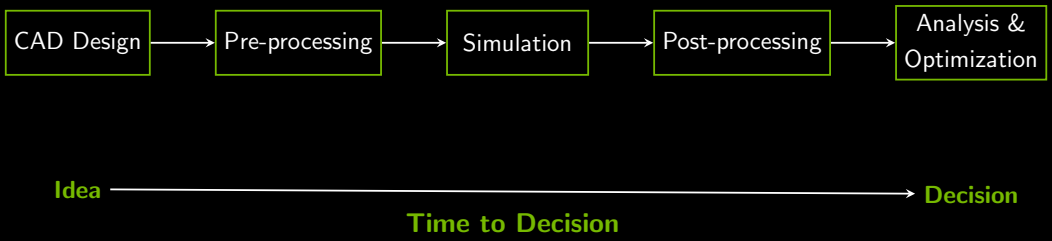
Future Work: The Storage Bottleneck

- HF simulations produce petabytes of data, outpacing HPC storage capacity.
- **Needs:** Advanced compression beyond simple subsampling.
- **Approaches:** Proper Orthogonal Decomposition (POD) and Latent-space compression via Autoencoders.

Future Work: Online Training

- **The In-Situ Paradigm:** Train ML model concurrently with the data-generating CFD simulation.
- **Advantage:** Eliminates need for intermediate storage and captures temporal dynamics inherently.
- **Challenge:** Managing vastly different timescales between physics generation and model backprop.

The Traditional CFD Workflow

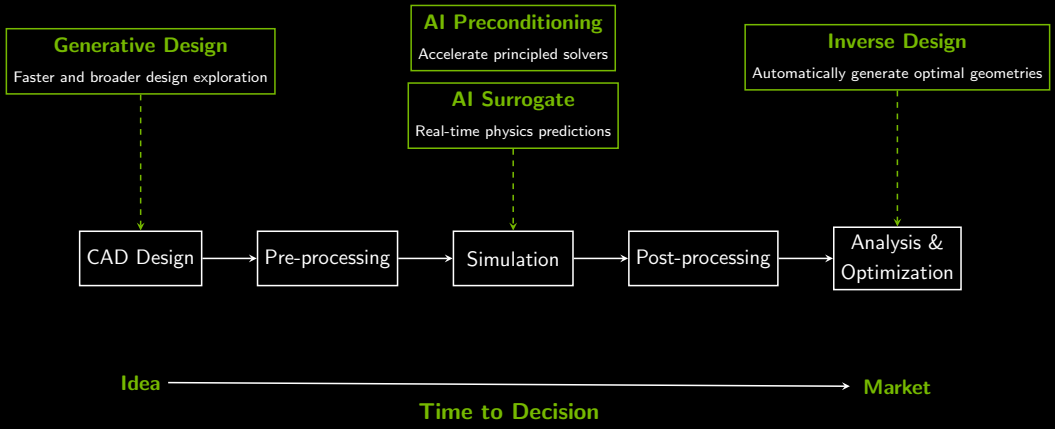


- The traditional engineering pipeline is highly serial and human-in-the-loop intensive.
- Progressing from Idea to Decision involves distinct, isolated steps.

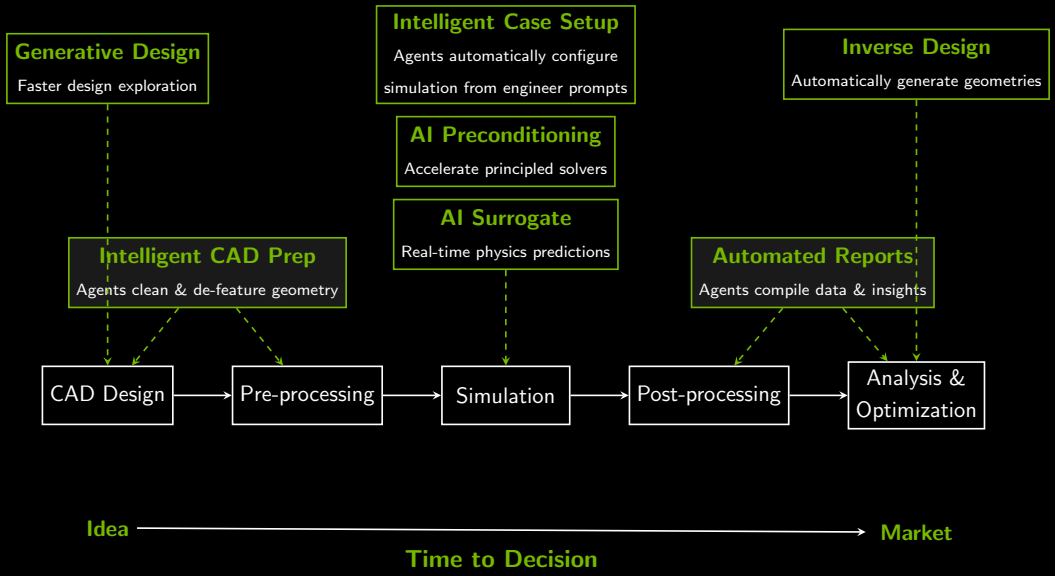
Amdahl's Law in Computational Engineering

- **Amdahl's Law:** $S = \frac{1}{(1-p)+\frac{p}{s}}$
- In CAE, p represents the parallelizable/accelerated portion (the Solver), while $(1-p)$ represents serial bottlenecks (CAD cleanup, Meshing, Post-processing).
- **The Bottleneck:** Even if we accelerate the solver by $100\times$ ($s = 100$), if CAD and meshing take 50% of the engineer's time ($(1-p) = 0.5$), the maximum theoretical workflow speedup is only $2\times$.
- To achieve true time-to-market reduction, we must accelerate **every part** of the pipeline, not just the simulation compute.

Engineering With AI: AI Surrogates and Design



Solving Amdahl's Law: Agentic Workflows



Enabling Agents: APIs and the Model Context Protocol

- For autonomous AI agents to manage CAD, meshing, and solvers, human-centric GUIs are no longer sufficient.
- **API-Driven Ecosystems:** CFD tools must expose robust APIs that allow AI orchestration layers to manipulate geometry, set boundary conditions, and extract data.
- **Interoperability Standards:**
  - Codebases must be built to be compatible with `skills.md` guidelines.
  - Adoption of the **Model Context Protocol (MCP)** allows foundation models to seamlessly connect to proprietary engineering data sources, solvers, and local tools.

Summary

- Building a CFD foundation model is a **tractable problem**.
- Scaling to the "universal" limit requires moving beyond RANS to high-fidelity transient data.
- Future progress is linked to **GPU-native CFD solvers** co-designed for ML framework integration.
- True acceleration mandates addressing Amdahl's Law holistically: coupling AI Surrogates for physics with **AI Agents** for workflow automation.