AA216/CME345: PROJECTION-BASED MODEL ORDER REDUCTION

Introduction

Charbel Farhat
Stanford University
cfarhat@stanford.edu

Outline

HEILMEIER CHATECHISM

- Critical questions for research proposal
  1. What is the problem? why is it hard?
  2. How is it solved today, what are the limits of current practice?
  3. What is the new technical idea? why can we succeed now?
  4. Who cares?
  5. What is the impact if successful?

Outline

BUFFETING: $M_\infty = 0.243$, AoA = 30.3^o

F/A-18 VERTICAL TAIL BUFFETING

F-18 BUILDER SAYS TEST MISSED DESIGN FLAWS
By Wayne Biddle, Special To the New York Times
July 28, 1984

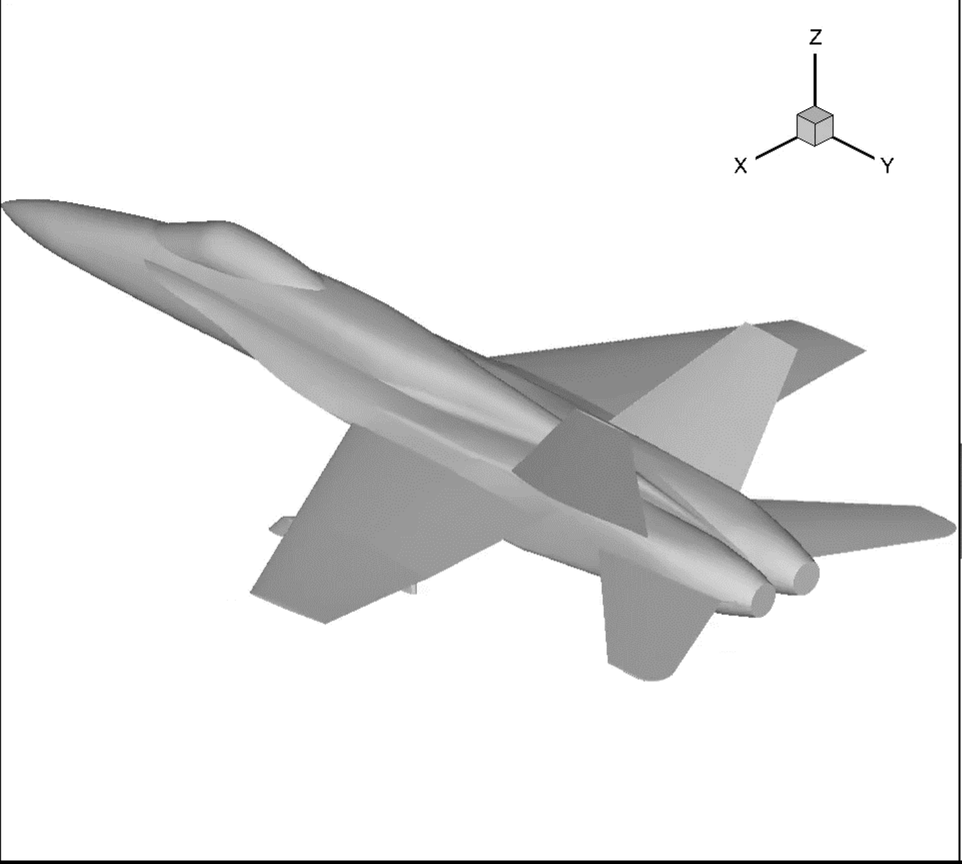
"Had we properly assessed it, we would have designed for it and not had the problems we have," Don Snyder, Director of F-18 engineering at the McDonnell Douglas Corporation (1984)

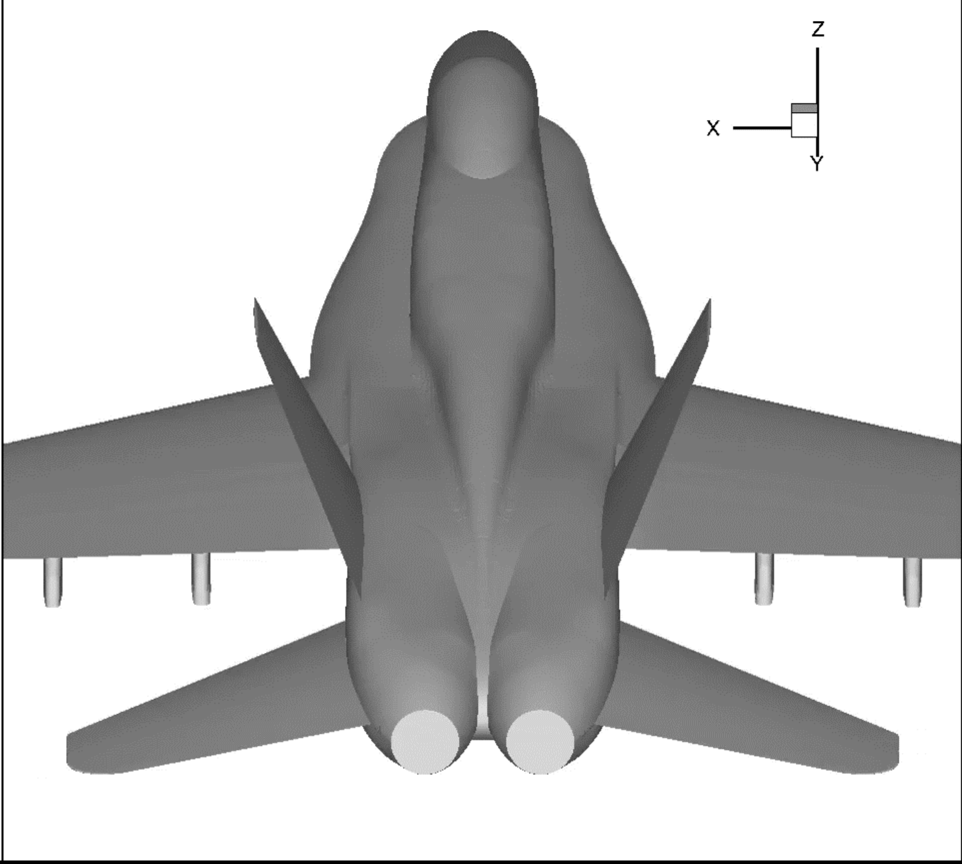
MCDONNELL DOUGLAS SAYS IT WILL PAY COST OF FIXING F-18 FLAW
By Wayne Biddle, Special To the New York Times
Aug. 3, 1984

The McDonnell Douglas Corporation announced today that it would *''bear the costs,''* about $25 million, of correcting a design flaw in the Navy's F-18 warplane that has caused cracks in the fighter's tail.

BUFFETING: $M_\infty = 0.243$, AoA = 30.3°

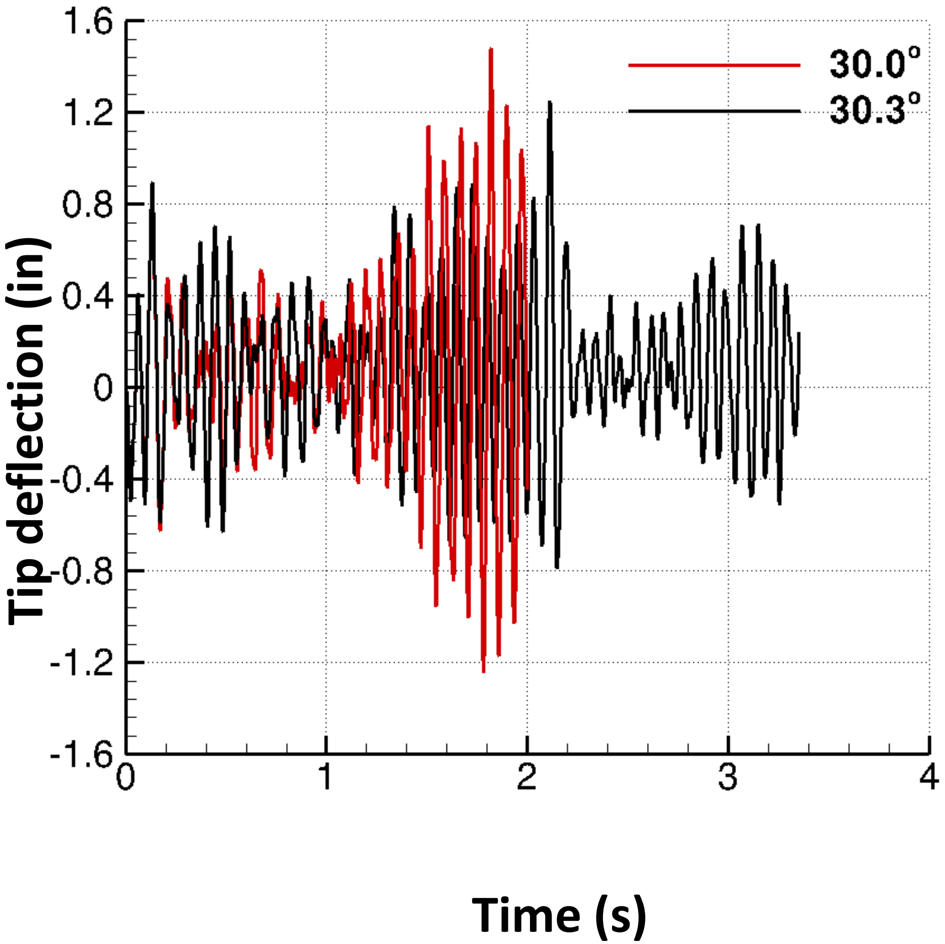
1,500 cores for solving a $450 M

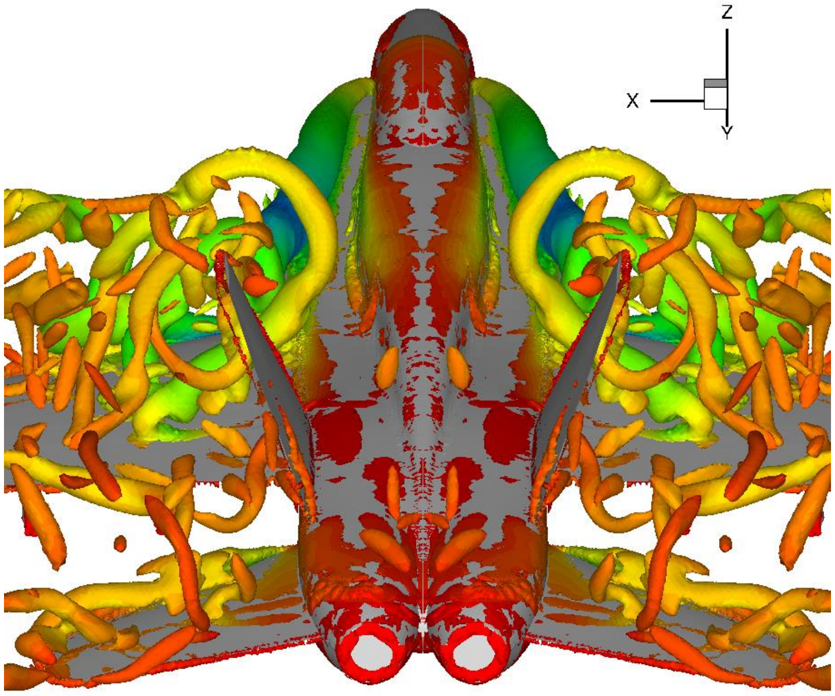




F/A-18 VERTICAL TAIL BUFFETING

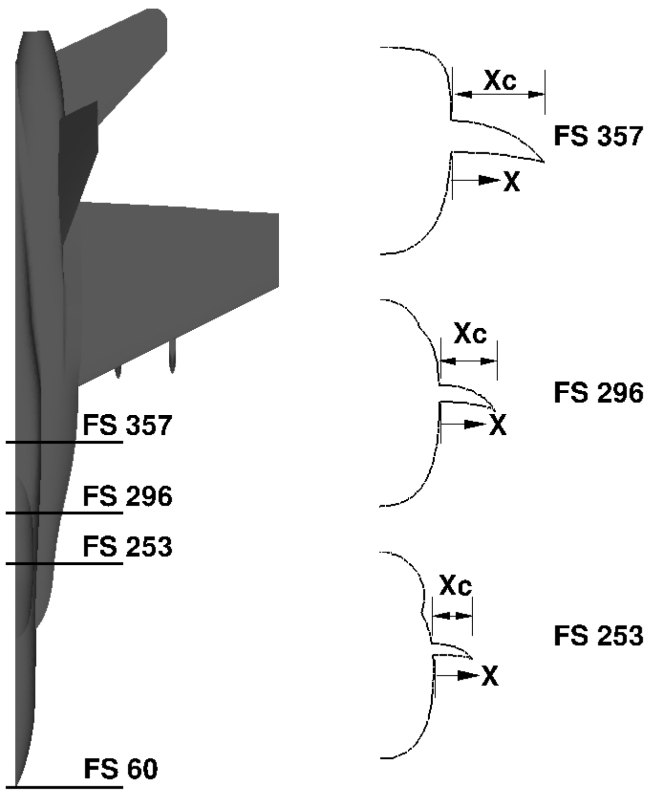
M_∞ = 0.243, AoA = 30.3°

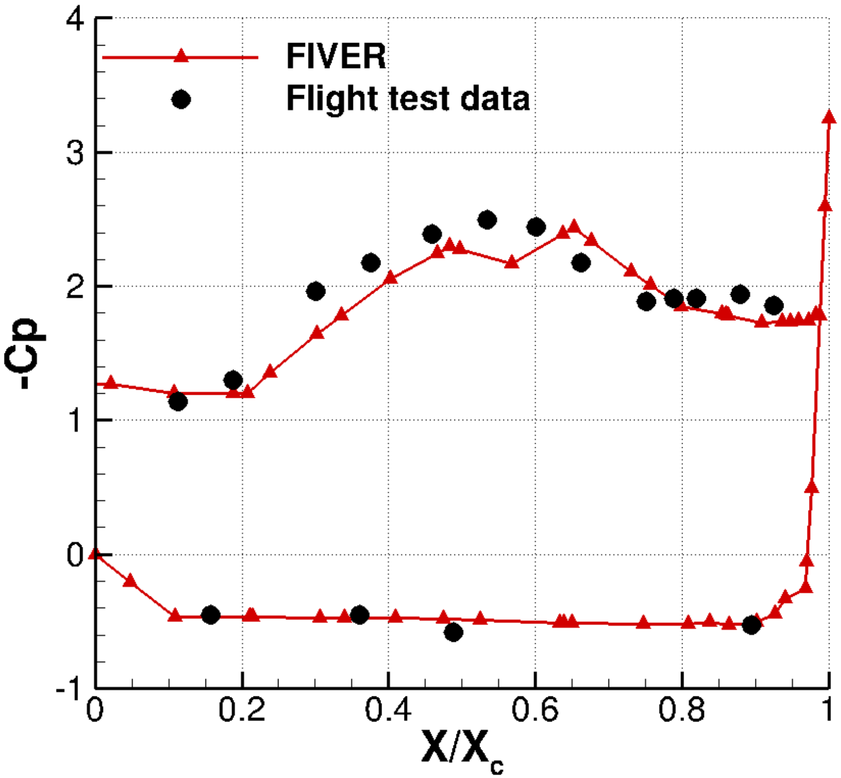




F/A-18 VERTICAL TAIL BUFFETING

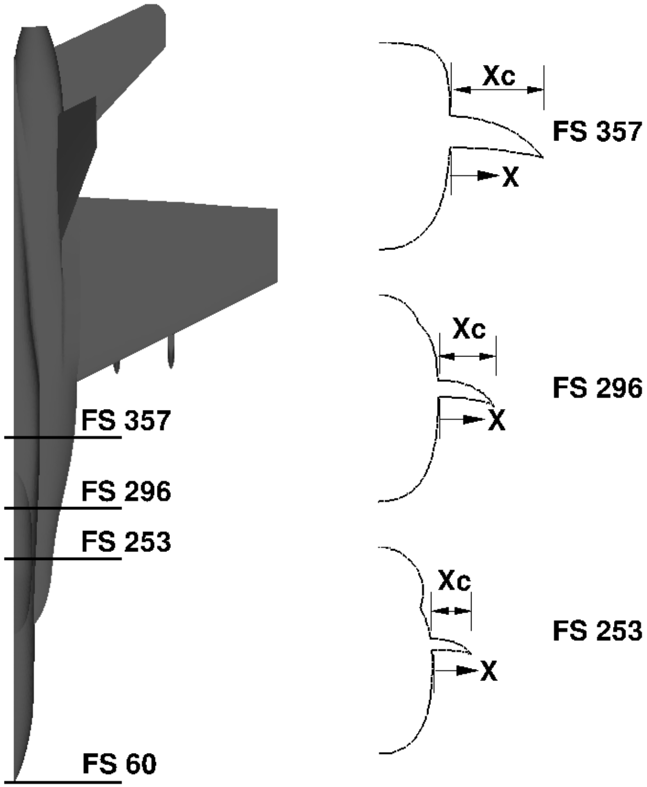
M_∞ = 0.243, AoA = 30.3°

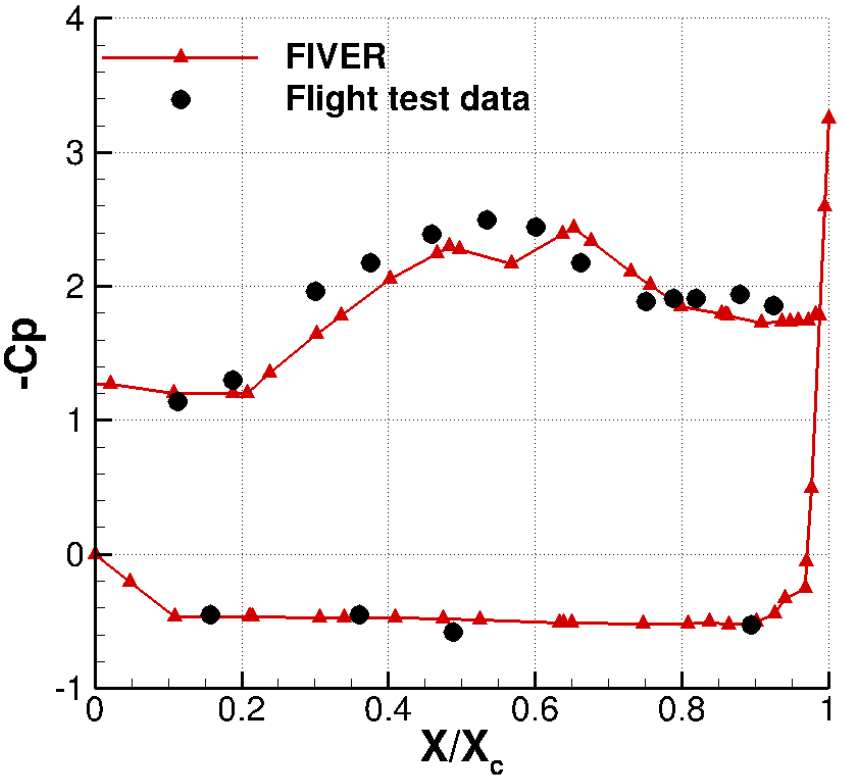




WHAT IS THE PROBLEM?

$M_\infty = 0.243$, AoA = 30.3º





4 days on a 1,500-core massively parallel system for a single configuration

WHAT IS THE PROBLEM?

- Parametric simulations
  - free-stream velocity
  - angle of attack
  - altitude
  - positions of the control surfaces
  - shape variations
  - structural design variations
  - …

→ 1,000s of simulations

HOW IS IT SOLVED TODAY? – LIMITS

- High performance (exascale) computing

- Impressive performance
  - 11 billion equations
  - 22,000 cores
  - 3 minutes wall-clock time

- Unimpressive performance
  - time-loop
    - o 50,000 time-steps
    - o 1 solve/time-step

  → ~ **<u>4 months wall-clock time</u>** …

- Time is one among many parameters of interest for engineering applications

Stanford Report, June 10, 2015

Stanford engineers team up with U.S. Army to set computational record

Now billions of questions can be answered in about three minutes.

BY JAMES URTON

Stanford engineers have partnered with the U.S. Army Research Laboratory (ARL) to set a computational record. Stanford Professor Charbel Farhat and his research team at the Army High Performance Computing Research Center (AHPCRC) used a new, high-end, massively parallel computer to demonstrate the power of algorithms that instruct processors to work together to solve challenging problems.

They directed 22,000 processors to solve billions of mathematical equations in just a few minutes, a rare feat in computer engineering.

"I believe we may have really set a record here," said Farhat, director of AHPCRC at Stanford and professor of aeronautics and astronautics and of mechanical engineering. "We've solved over 10 billion equations in a little over three minutes.

This unique collaboration came about after the ARL acquired Excalibur, a Cray XC40 computer with 101,184 processors. Once completely up and running at ARL's Maryland-based Defense Computing Resource Center (ARL DSRC), the raw computational power of this system will support dozens – potentially hundreds – of research

Norbert von der Groeben

Stanford Professor Charbel Farhat and his team accomplished a rare feat in computer engineering through a partnership with the U.S. Army Research Laboratory.

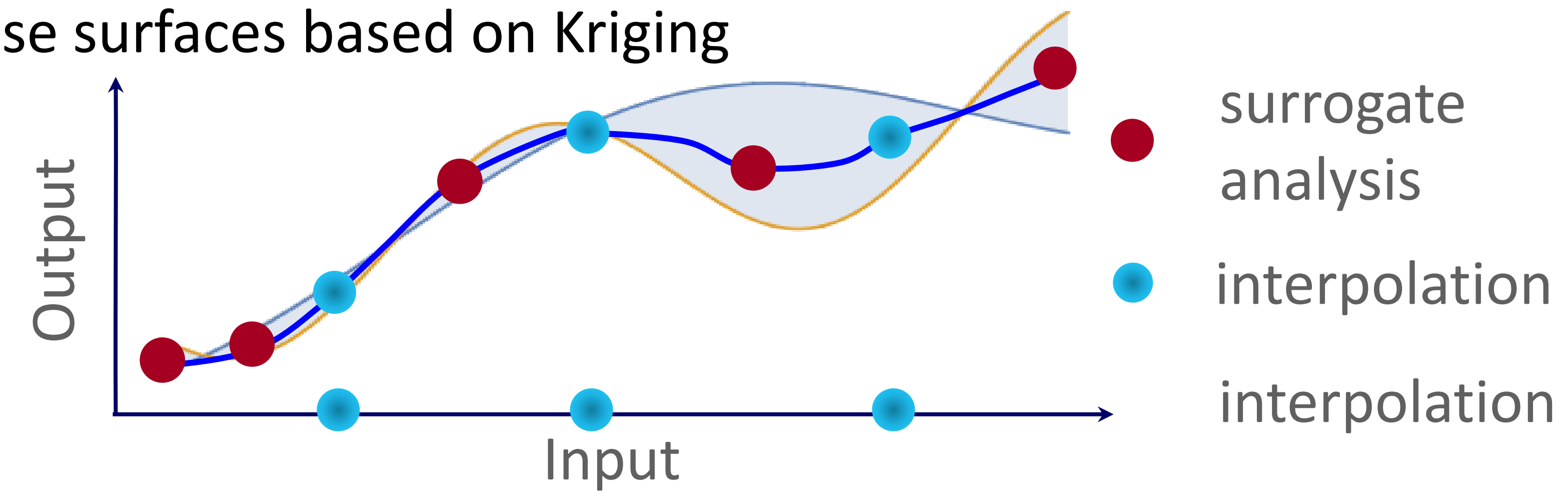
HOW IS IT SOLVED TODAY? – LIMITS

- Surrogate models

input → SURROGATE MODEL OF A SYSTEM → output

$\mu = [\mu_1,\ldots,\mu_p]$ → SURROGATE MODEL OF A SYSTEM → $y = [y_1,\ldots,y_p]$

- Response surfaces based on Kriging



- method of interpolation for which the interpolated values are modeled by a Gaussian process governed by prior covariances
- a model of the ***output*** but not of the system
- limits: ***few***, ***pre-determined, scalar*** quantities of interest (QoIs)

Outline

2 What Is Model Reduction?

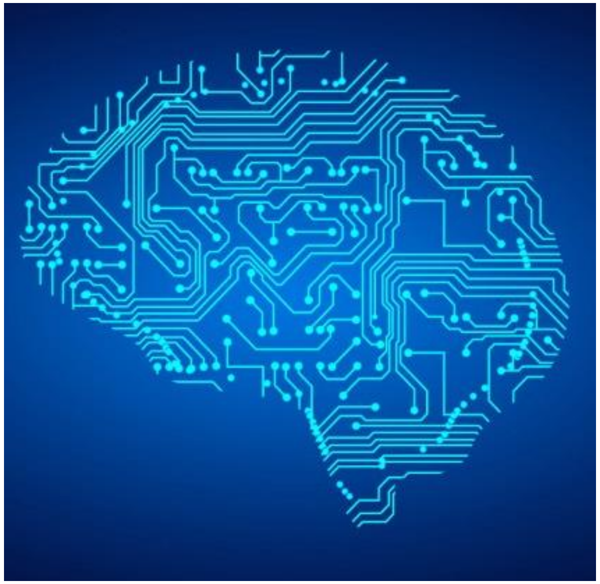
WHAT IS THE NEW TECHNICAL IDEA?

- Surrogate models

input $\longrightarrow$ SURROGATE MODEL OF A SYSTEM $\longrightarrow$ output

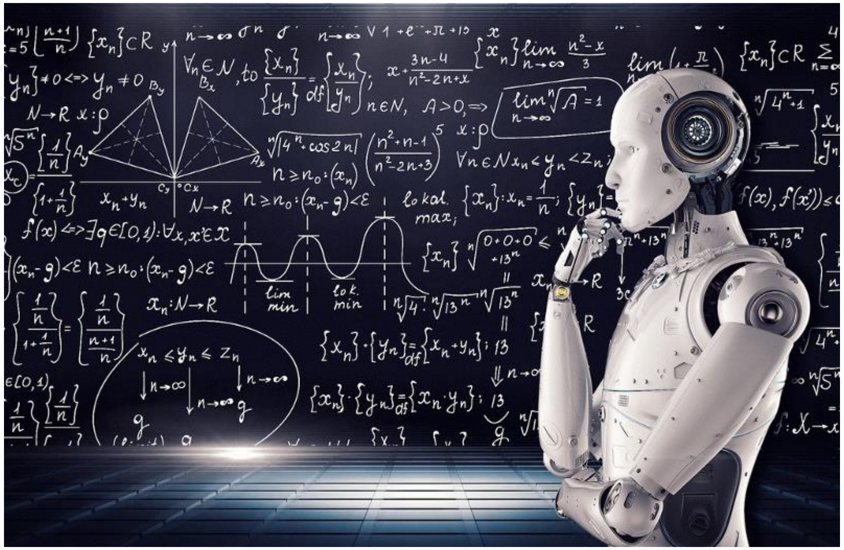
$\mu = [\mu_1, \ldots, \mu_p]$ $\qquad$ $y = [y_1, \ldots, y_p]$

- Regression artificial neural networks
  - universal approximation theorem
  - very large amount of training data
  - limits: ***few***, ***pre-determined, scalar*** QoIs
    few, medium-size, vector QoIs



WHAT IS THE NEW TECHNICAL IDEA?

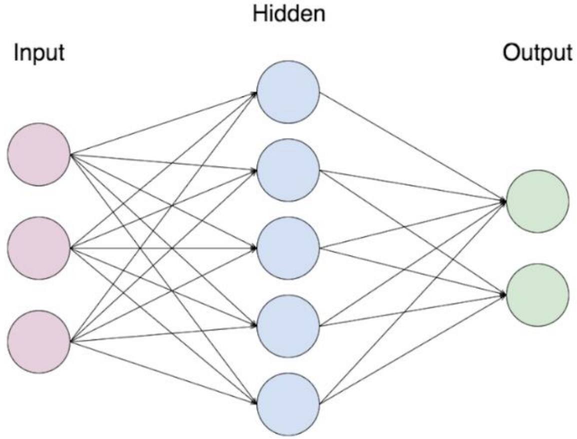
- Machine Learning (ML)



- *classical* ML algorithms use computational methods to "learn" information directly from data ***without relying on a predetermined equation as a model***
- *classical* ML algorithms have been mostly driven by problems for which ***no predetermined equation is available*** → data-driven modeling

WHAT IS THE NEW TECHNICAL IDEA?

- Reading comprehension, speech recognition, face recognition, …
  - large amounts of data are available and are free of charge to the analyst
  - the concepts of error and accuracy are application-dependent
  - the concept of prediction is more qualitative than quantitative



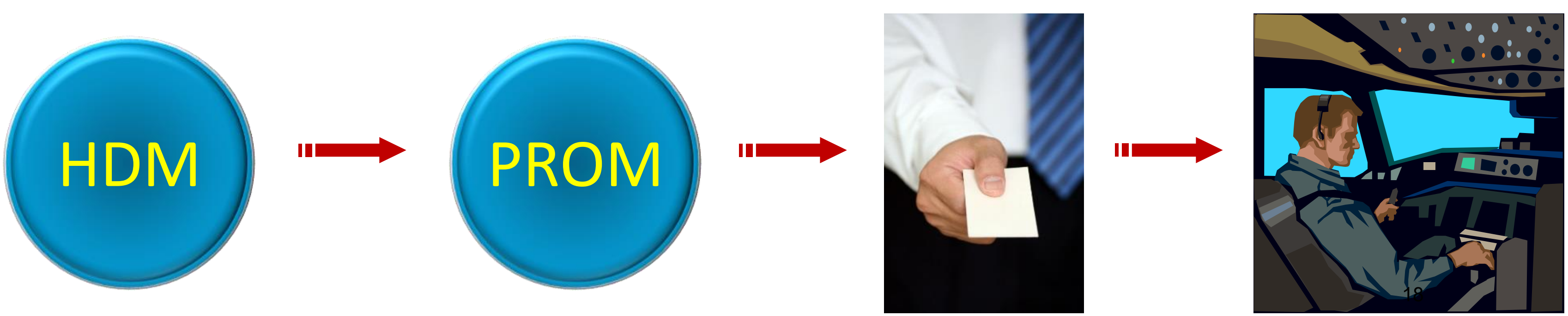
- Physical systems
  - little amounts of data are available and are difficult/expensive to obtain
  - the concepts of error and accuracy are universal
  - the concept of prediction is mostly quantitative

WHAT IS THE NEW TECHNICAL IDEA?

- Projection-Based Model Order Reduction (PMOR)
  - build the ***lowest-dimensional model*** that can capture the dominant behavior of the system of interest by ***projecting a given High-Dimensional computational Model (HDM) on a subspace constructed after learning something about the system of interest***

→ Projection-Based Reduced-Order Model (PROM)

→ compact representation and drastic CPU time reduction at minimum loss of fidelity



WHY IS THIS PROBLEM HARD?

- Can the solution of every problem of interest be approximated in a low-dimensional space?
  - at what cost?
- The curse of parameterization
  - how to achieve robustness with respect to parameter changes?
  - how to mitigate the effect of the dimensionality of the parameter space on that of the solution space?
- Similar difficulties as in the case of HDMs
  - numerical stability, accuracy, boundary conditions, arbitrary constraints, error estimators, …
- Compact representation?
  - separation of reduced bases and generalized coordinates
  - intrusiveness
- What logistics are needed to recover the investment?

COURSE SCOPE: LINEAR PMOR

- PMOR
  - physics-based model → Partial Differential Equation (PDE)
  - linear(ized) → semi-discrete level
  - nonlinear → discrete level

- Linearized (or linear) problems
  - circuits
  - acoustics (frequency domain)
  - structural dynamics
  - stability
  - control
  - sensitivity

- Mature for zero- and one-parameter problems (i.e., frequency)
  - textbooks: [Antoulas, 2010], [Benner, 2011]
  - commercial software (ANSYS)

Outline

3 How Is It Done Today?

PMOR = PHYSICS-BASED ML

- Nonlinear, parametric, dynamical system (implicit discrete level)

$$\mathbf{r}(\mathbf{u};\boldsymbol{\mu}) = 0, \quad \mathbf{u} = \mathbf{u}(t^m), \quad \mathbf{r} \in \mathbb{R}^N$$

- Hypothesis: affine approximation in some region of the state-space

$$\mathbf{u} \approx \mathbf{u}_0 + \mathbf{V}\Delta\mathbf{y}, \quad \mathbf{V} \in \mathbb{R}^{N\times n}, \quad \Delta\mathbf{y} \in \mathbb{R}^n, \quad n \ll N$$

- Representation: reduced-order basis (ROB) $\mathbf{V}$

- Data-driven learning process: $\mathbf{V}$ is learned from data generated by exercising a high-dimensional model (HDM) at points sampled in a parameter space using a greedy procedure, computing solution snapshots & compressing them using SVD

- Loss function and optimization problem

$$\mathbf{u} \approx \mathbf{u}_0 + \mathbf{V}\Delta\mathbf{y} \rightarrow \mathbf{r}(\mathbf{u}_0 + \mathbf{V}\Delta\mathbf{y};\boldsymbol{\mu}) \approx \mathbf{r}(\mathbf{u}_0;\boldsymbol{\mu}) + \mathbf{J}(\mathbf{u}_0;\boldsymbol{\mu})\mathbf{V}\Delta\mathbf{y} \approx 0$$

$$\boxed{\Delta\mathbf{y} = \underset{\mathbf{z}\in\mathbb{R}^n}{\arg\min} \|\mathbf{J}^{-1}\mathbf{r}(\mathbf{u}_0;\boldsymbol{\mu}) + \mathbf{V}\mathbf{z}\|_{\Theta}}$$

physics-based machine learning approach for constructing a lower-dimensional, structure-preserving model

OFFLINE-ONLINE COMPUTING

HDM (μ^1) | HDM (μ^2) | ⋮ | HDM (μ^l) | PROM (μ^{l+1}) | PROM (μ^{l+2}) | PROM (μ^{l+3}) | PROM (μ^{l+4}) | ⋮ | PROM (μ^m)

snapshots

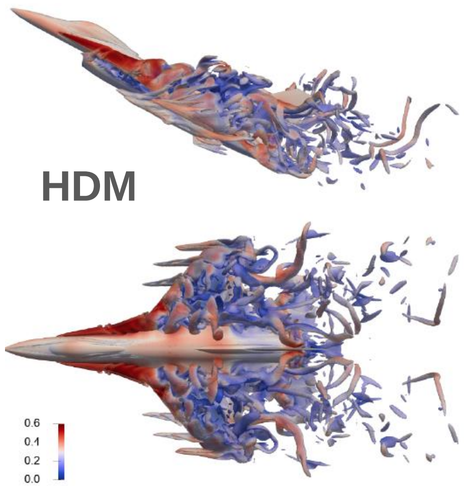
compress (SVD)

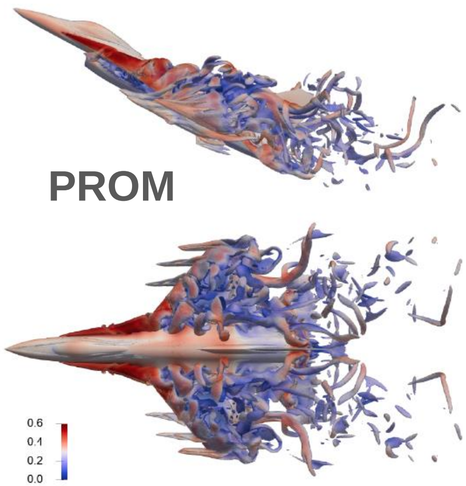
ROBs V_i

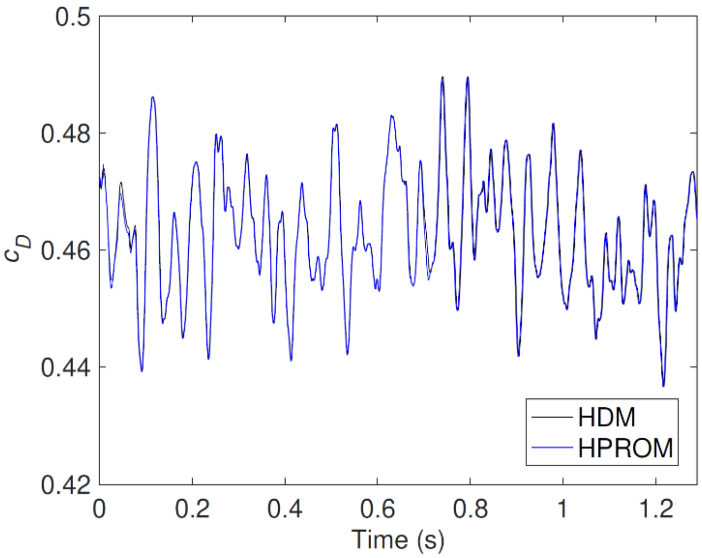
projection-based parametric PROM(s)

STATE OF THE ART

- F-16 C/D Block 40 at M_∞ = 0.3, 30° angle of attack, Re = 18,200,000
  - ***DES***; CFD mesh with 26,919,879 vertices and ***158,954,429 tetrahedra***
  - time-interval $t \in [0, 1.29]$ s



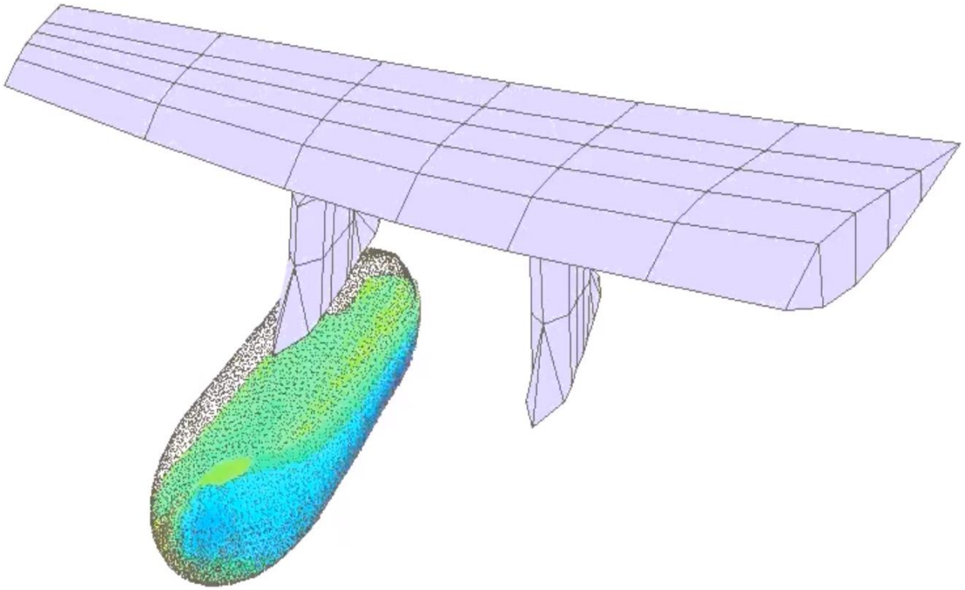


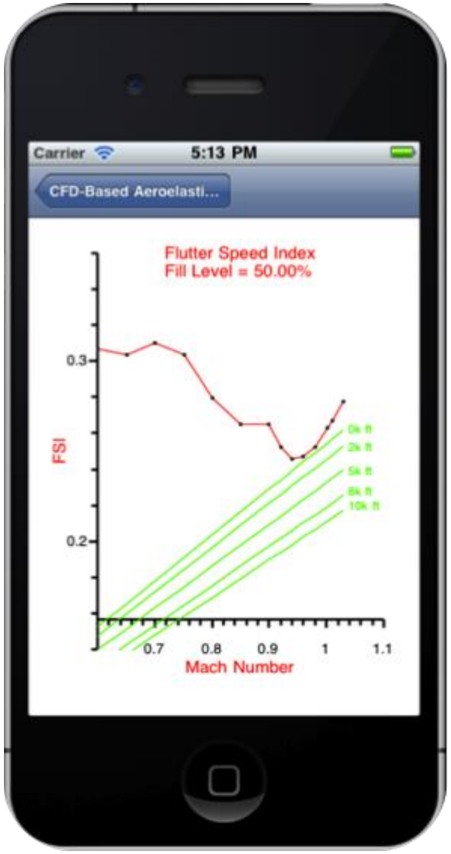


- ***Frontera*** supercomputer at the University of Texas at Austin
  - o HDM: ***100.3 hrs wall-clock time*** on ***3,584 cores*** (per configuration)
  - o PROM: ***5.8 min wall-clock time*** on ***32 cores***
  - o ***speedup factor*** (wall-clock time) = ***1.04 × 10***^{3}
  - o ***speedup factor*** (CPU time) = ***1.17 × 10***^{5}

LEARNING WITH MODELS AND DATA vs. LEARNING WITH DATA ONLY

- Real-time prediction of fuel sloshing and its effect on flutter

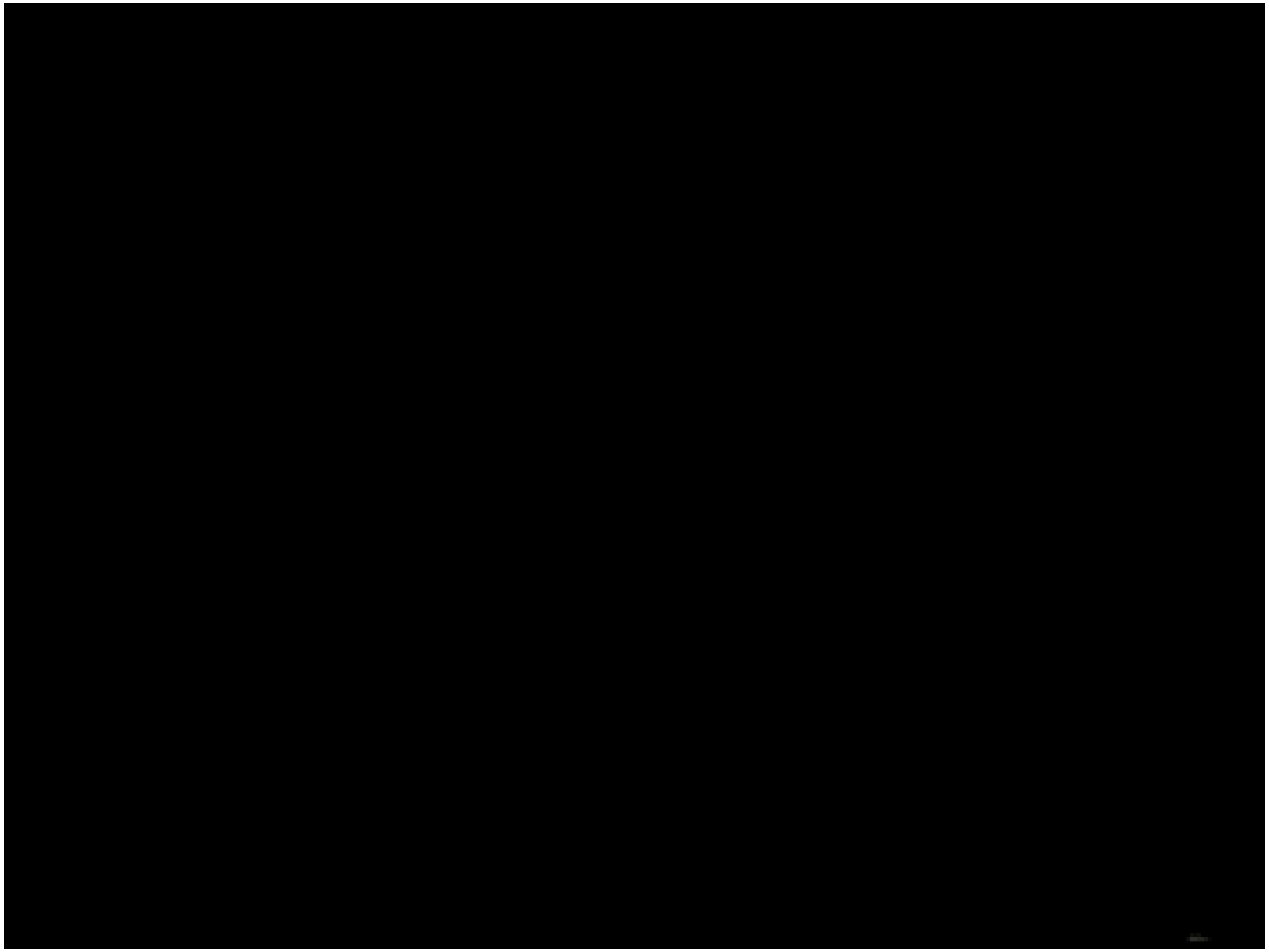




LEARNING WITH MODELS AND DATA vs. LEARNING WITH DATA ONLY

"Some fear flutter because they don't understand it, and some fear it because they do."

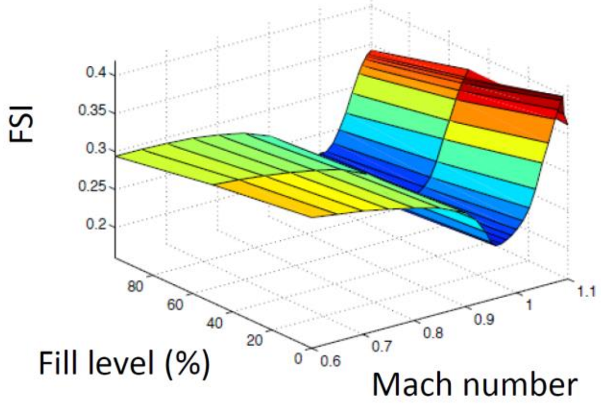
-von Karman-



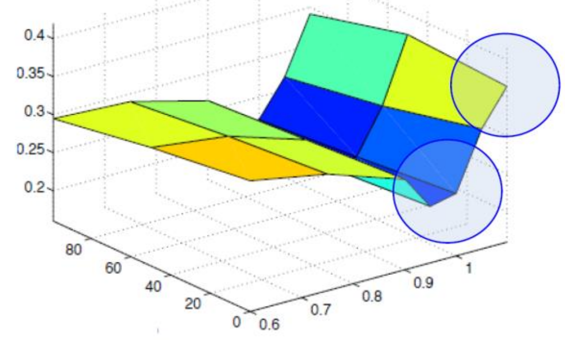
LEARNING WITH MODELS AND DATA vs. LEARNING WITH DATA ONLY

- Learning the flutter speed index (FSI)

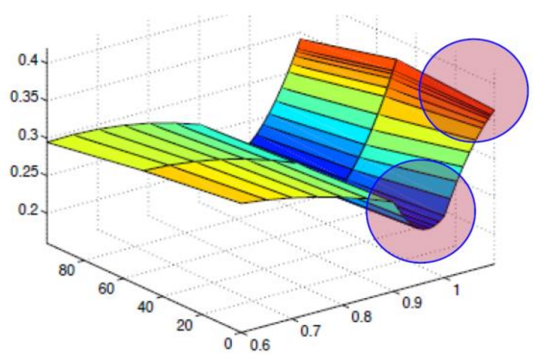
Ground truth data



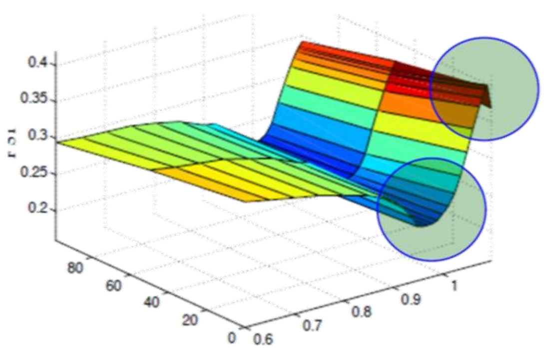
Database



Data regression (cubic)

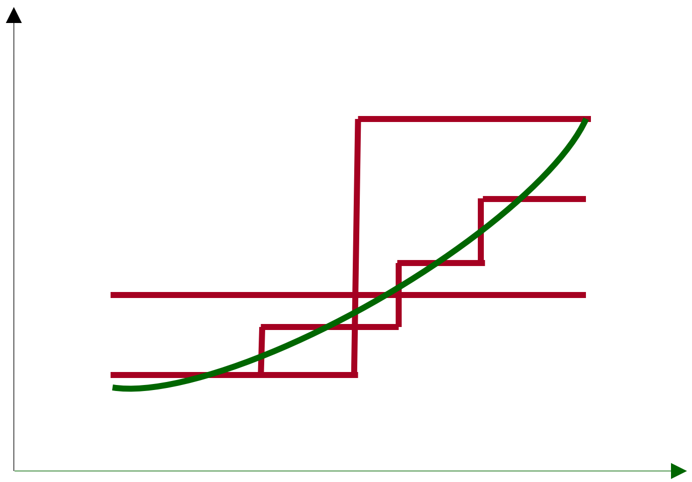


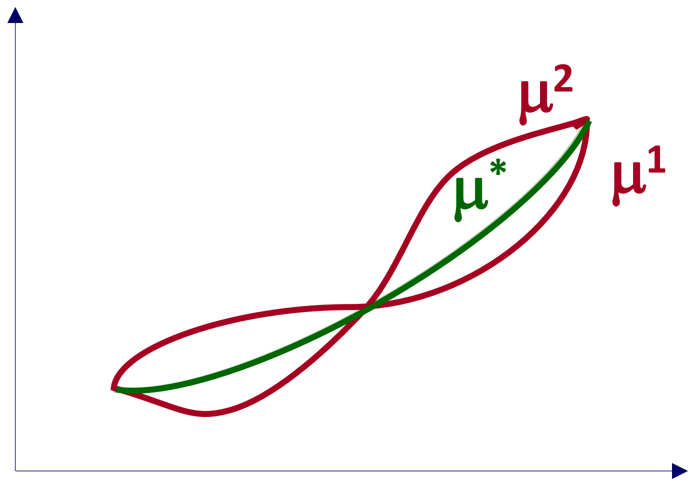
Model regression (linear)



WHAT IS THE NEW TECHNICAL IDEA?

- High-dimensional computational models are constructed *a priori*, using *local* polynomial basis functions which do not know much about the dynamical system of interest



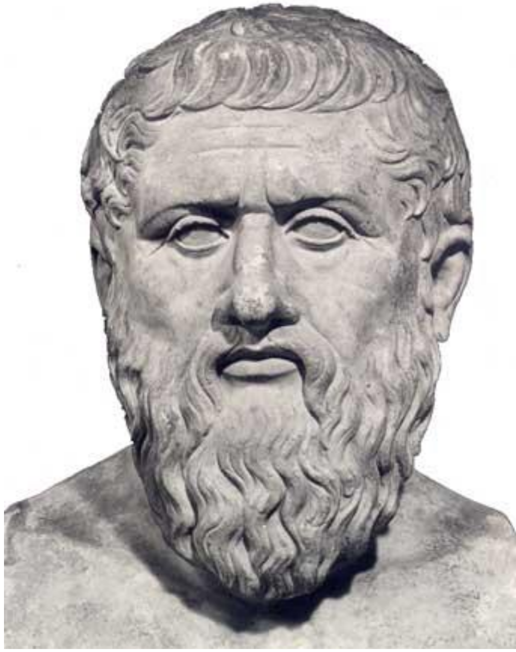


- Parametric PROMs are constructed using *global* basis functions which embody *a posteriori* information gathered from exercising the HDM offline to compute solution snapshots

$$\mathbf{u} \approx \mathbf{u}_0 + \mathbf{V}\Delta\mathbf{y}, \quad \mathbf{V} \in \mathbb{R}^{N \times n}, \quad \Delta\mathbf{y} \in \mathbb{R}^n, \quad n \ll N$$

WHAT IS THE NEW TECHNICAL IDEA?

- "We do not learn new things, we merely remember things we have forgotten"



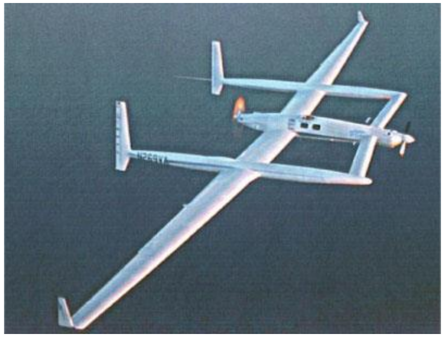
Plato

- PMOR is a Ritz method where the global basis functions are constructed a posteriori after some knowledge about the parameterized system has been developed, instead of being selected a priori

- Challenges
  - satisfaction of the parameteric boundary conditions by the global ROB V
  - satisfaction of other parametric constraints
  - stability conditions where applicable (*many types*)

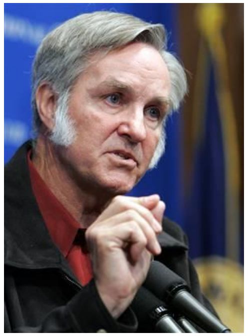
WHO CARES?

- Compute-intensive science and applications
  - parametric studies, stochastic analysis, uncertainty analysis
  - multidisciplinary modeling, multiscale modeling
  - multidisciplinary design optimization, optimal control, …



" [If I am not getting the NASTRAN answer after 4 hours on a Cray, then God is sending me the message I have the wrong design] "

Burt Rutan, 1993



- Time-critical applications (technology & industrial representatives)
  - embedded systems, virtual reality, robotic surgery
  - Boeing, Intel, Toyota, VW, ANSYS, ESI, ..
- Funding agencies (omnipresent in most recent initiatives)
  - combustion, hypersonics (DOD)
  - digital twins (DOE)

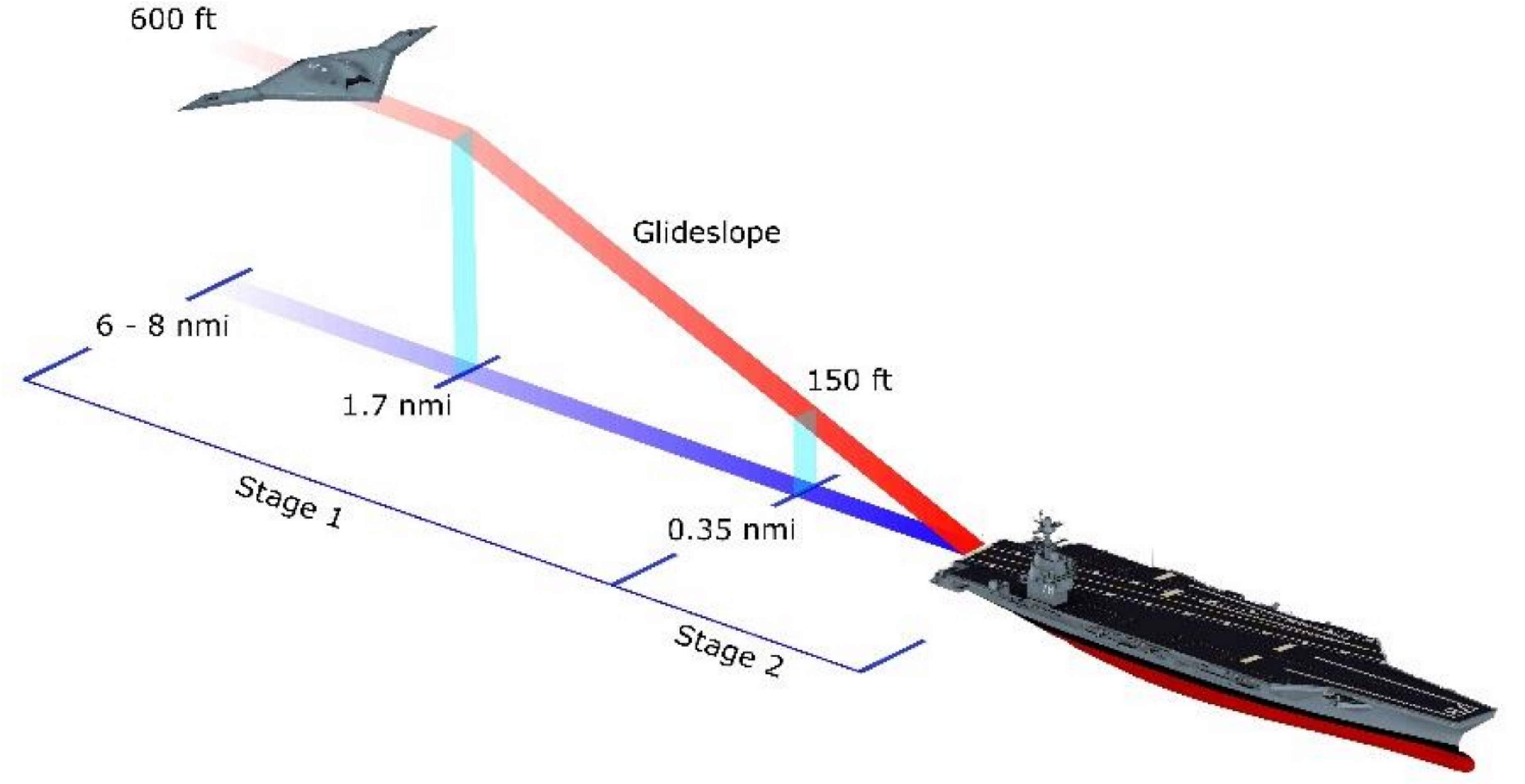
WHAT IF THE IMPACT IS SUCCESSFUL?

Adapted from "X-47B UCAS Aviation History Under Way" by Northrop Grumman,
Retrieved from https://www.youtube.com/watch?v=WC8U5_4lo2c

WHAT IF THE IMPACT IS SUCCESSFUL?

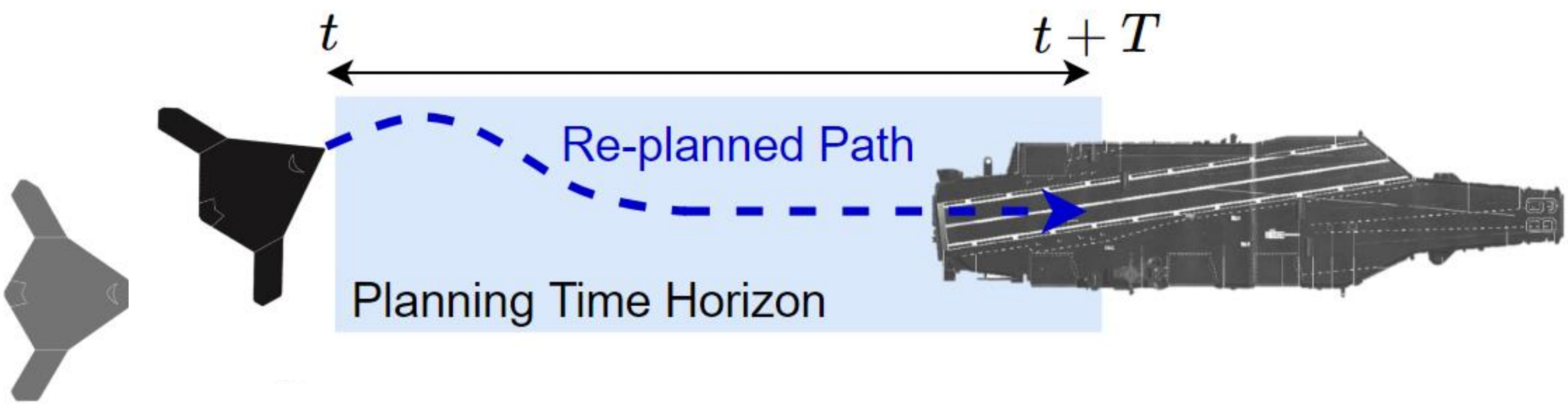
WHAT IF THE IMPACT IS SUCCESSFUL?

- Autonomous systems – Model Predictive Control (MPC)



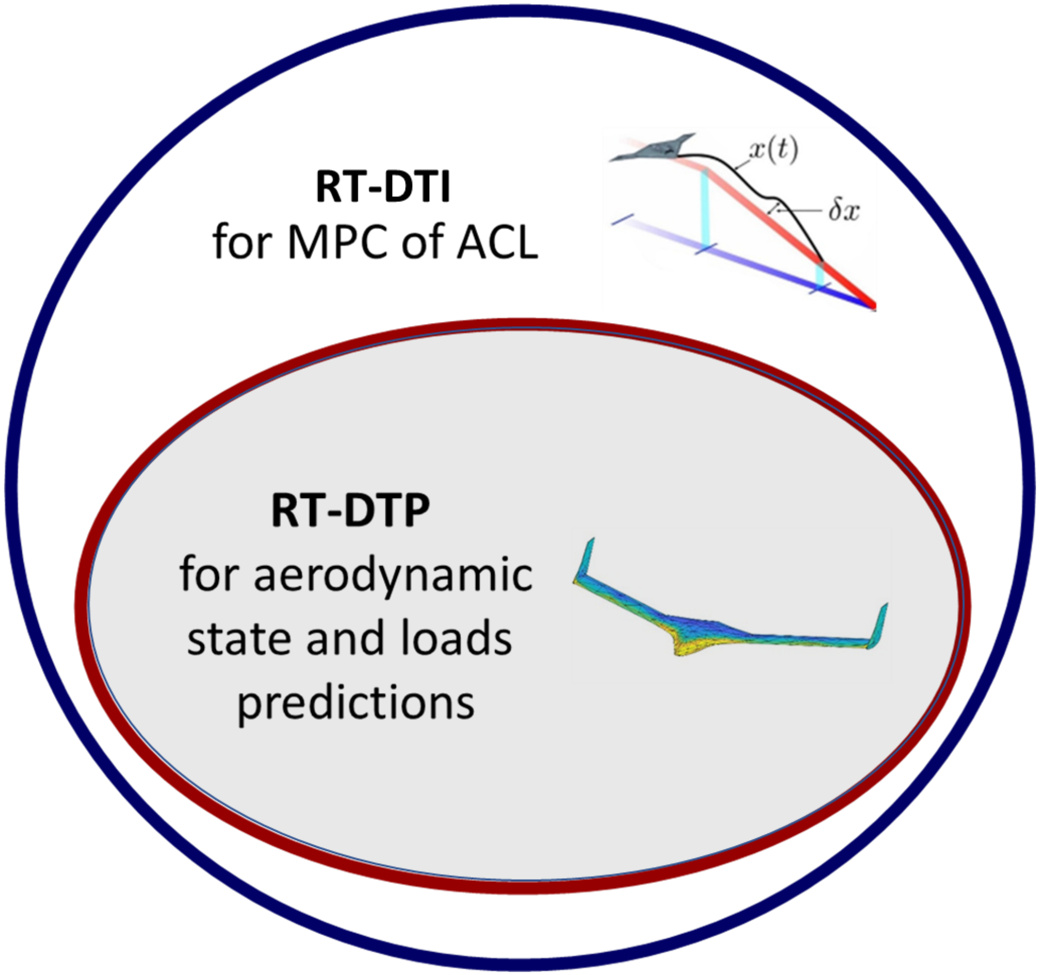
WHAT IF THE IMPACT IS SUCCESSFUL?

- Autonomous systems – Model Predictive Control (MPC)
  - principled method in optimal control theory
  - utilizes a computational model to optimize a system and predict its future behavior (OCP)
  - leverages state measurements to incorporate feedback into the system
  - accounts for state and control constraints and therefore may enable autonomous abort, and operation at performance limits



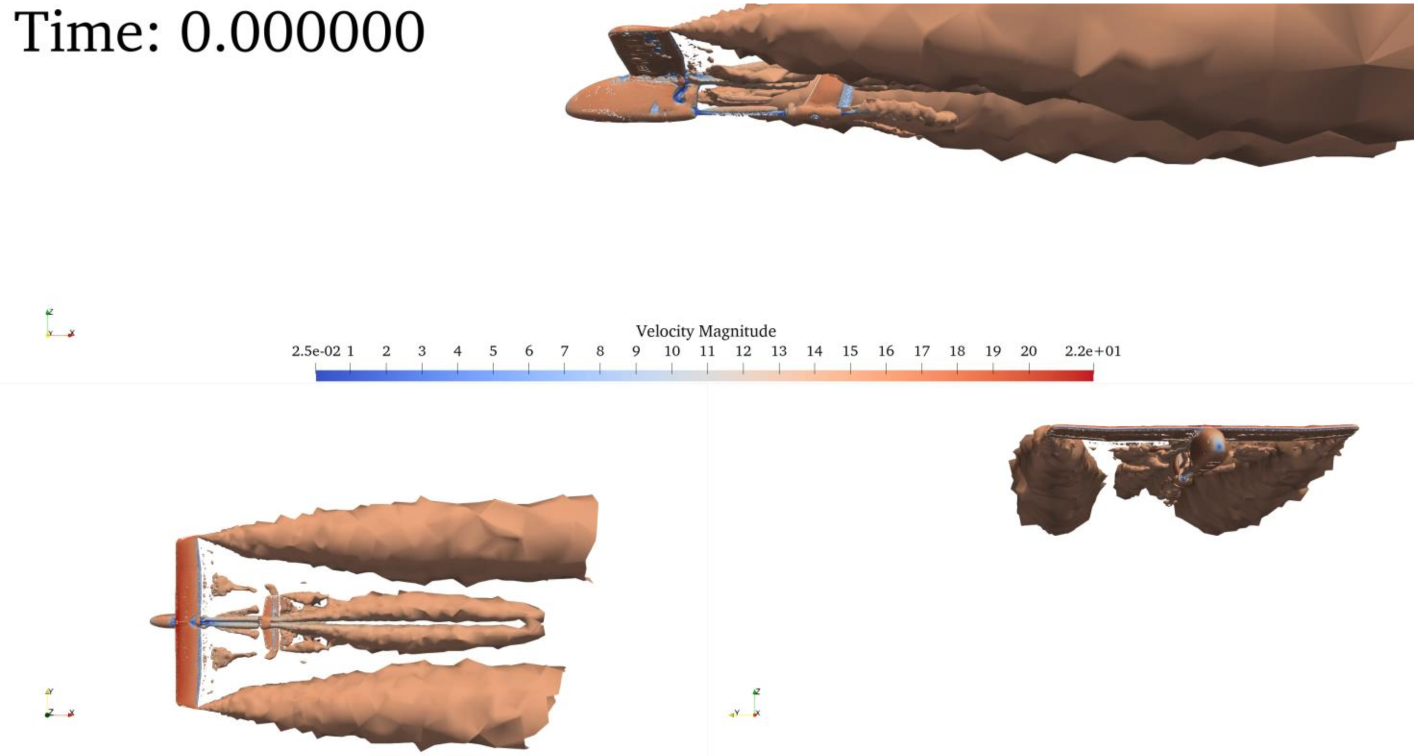
WHAT IF THE IMPACT IS SUCCESSFUL?

- PROM-based two-level digital twin



WHAT IF THE IMPACT IS SUCCESSFUL?

- Line-up error corrections and disturbance eliminations



RT-DTI	Wall clock time (msecs)	# of calls
state-estimator	6.7×10^{-2}	2,490
OCP	8.4	250

WHAT IF THE IMPACT IS SUCCESSFUL?

- Embedded RT-DTI (with Prof. M. Pavone)

WHY CAN WE SUCCEED NOW?

- Advances in approximation theory
  - error bounds for PROMs
  - Gaussian Processes (GPs)
  - …
- Emerging fields
  - machine learning (autoencoders)
  - data analytics
  - data analytics
- Advances in signal processing algorithms

Outline

4. When Does It Pay Off?

FEASIBILITY

- Whenever the cost of the offline phase can be amortized
  - parametric studies
  - stochastic analysis
  - uncertainty analysis
  - multiscale modeling
  - optimal control
  - design optimization

  .

  .

  .

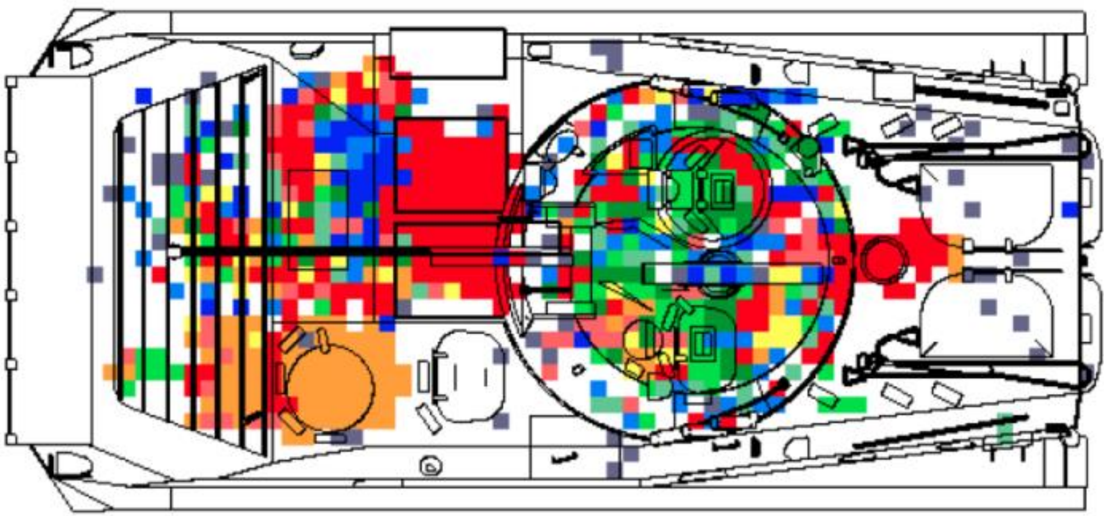
  - optimization of offline step cannot be ignored

  - particular attention to the sampling of a high-dimensional parameter space

  - data organization and manipulation

Outline

UNDER BODY BLAST



- too many parametric configurations to test or simulate
    - o charge intensity
    - o charge depth
    - o charge location
    - o standoff distance as a functions of armor parameters

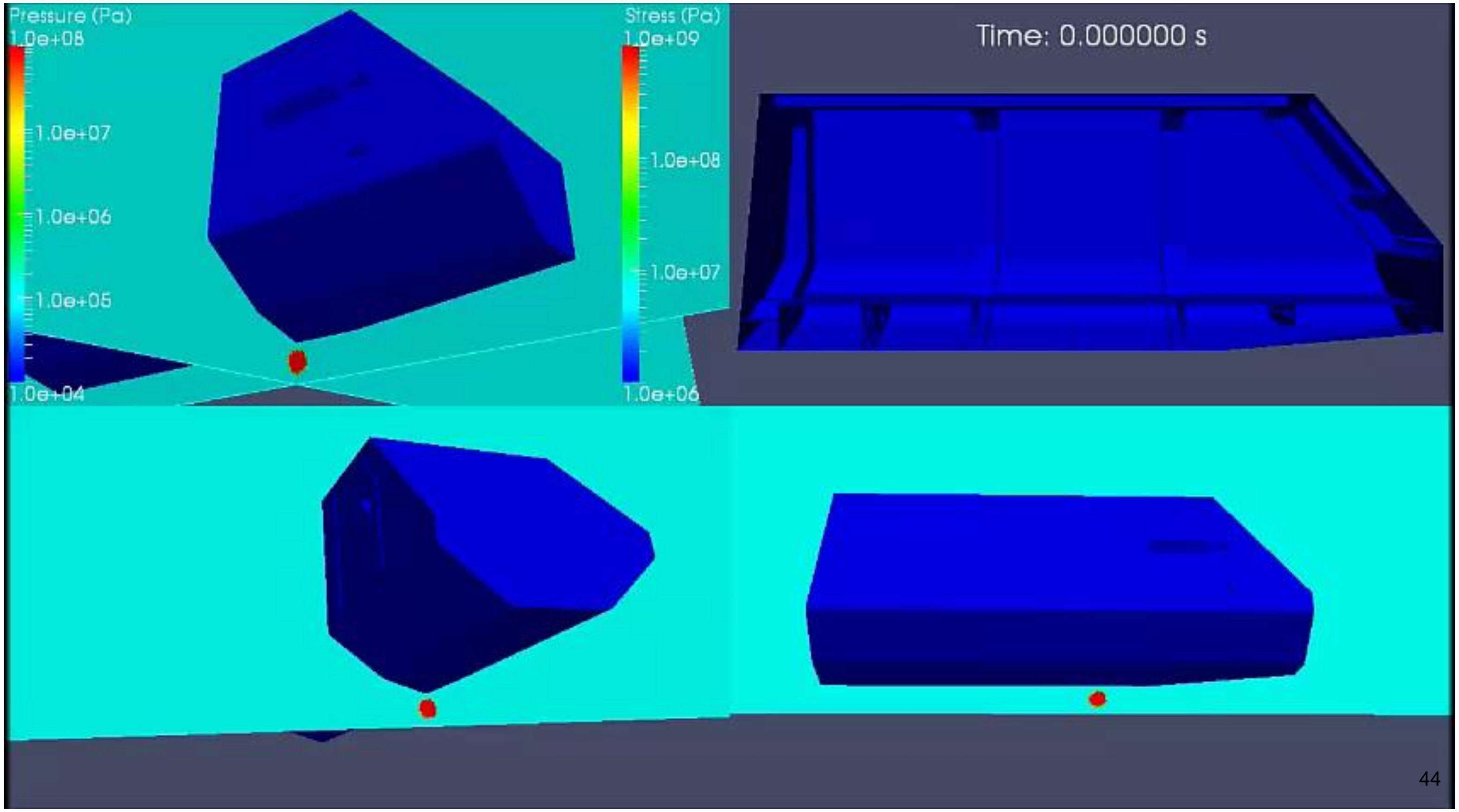
GENERIC V-HULL - HDM

- Nonlinear finite element model
  - J2 plasticity constitutive model
  - shells & beams (finite rotations)
  - 233,276 nodes
  - 236,995 elements
  - 1,399,056 dofs

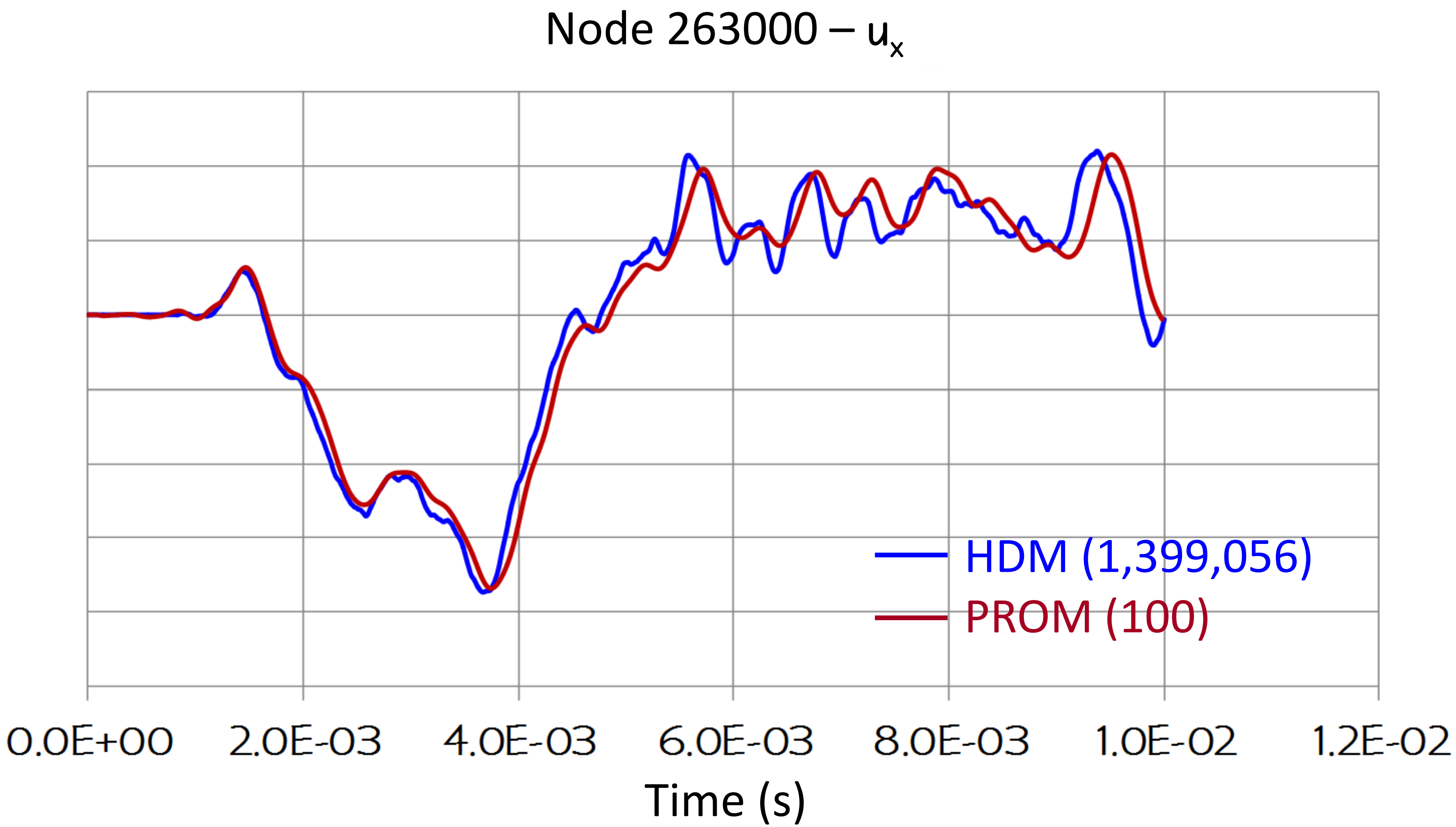
- Air blast wave
  - CONWEP module (10 kg charge)
- Explicit transient dynamic analysis
- POD-based PROM of dimension $n = 100$

GENERIC V-HULL - HDM

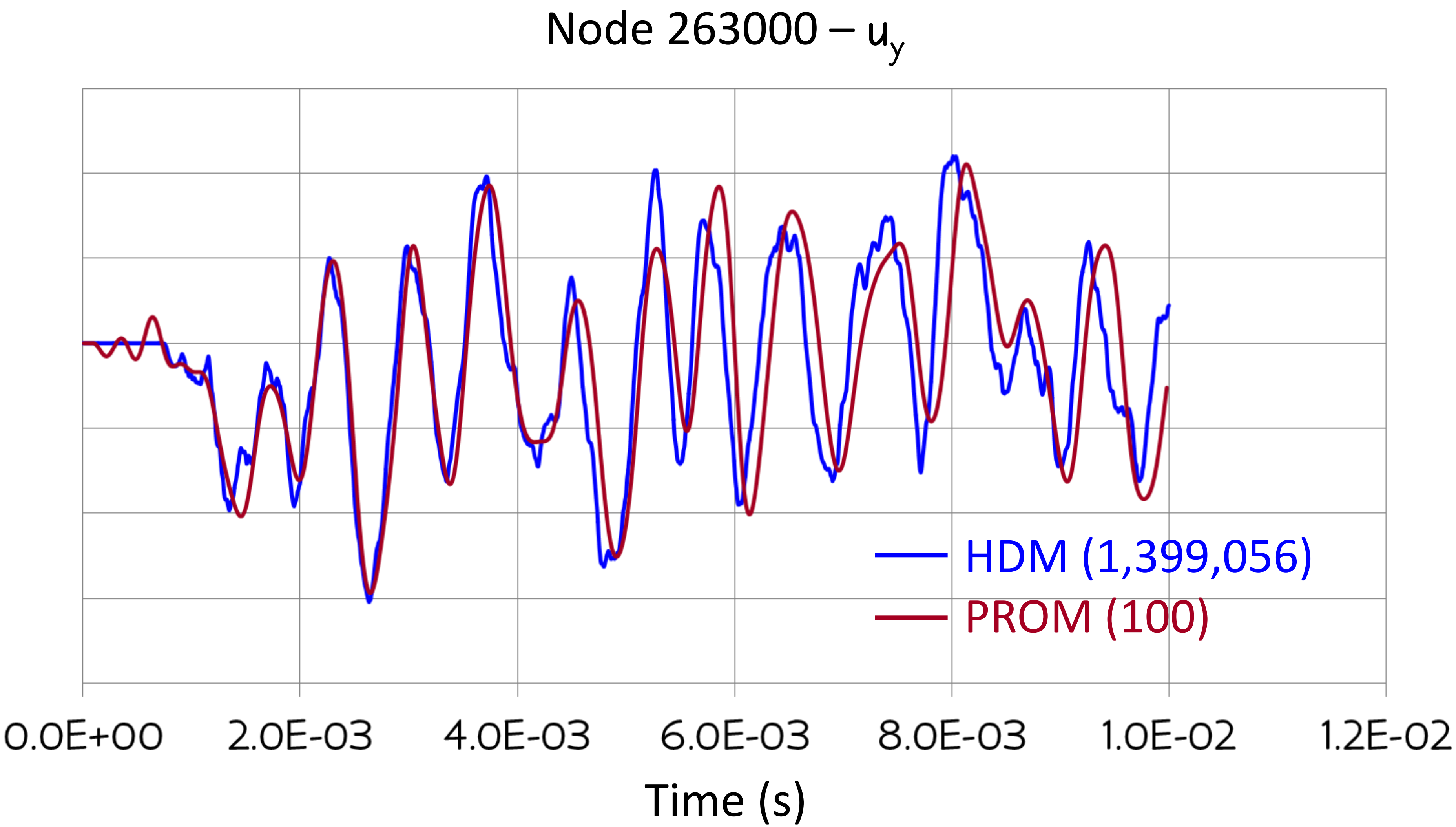
- Nonlinear structural dynamic response to air blast loading
  - simulation time-interval: $[0, 10^{-3}]$ s
  - midpoint rule: $\Delta t_{HDM} = 1 \times 10^{-8}$ s $(\Delta t_{PROM} = 2.5 \times 10^{-6}$ s$)$



PROM PERFORMANCE: ACCURACY



PROM PERFORMANCE: ACCURACY



PROM PERFORMANCE: SPEED-UP

- HDM: 128 cores of a Linux cluster — PROM: 16 cores

Model	Wall clock time	Speedup
HDM (1,399,056)	1.25×10^5 s (34.72 hrs)	
PROM (100)	4.32 s	28,935

→ CPU time speed-up = 231,481

VISION FOR THE FUTURE

- Next-generation computing-testing using PROMs

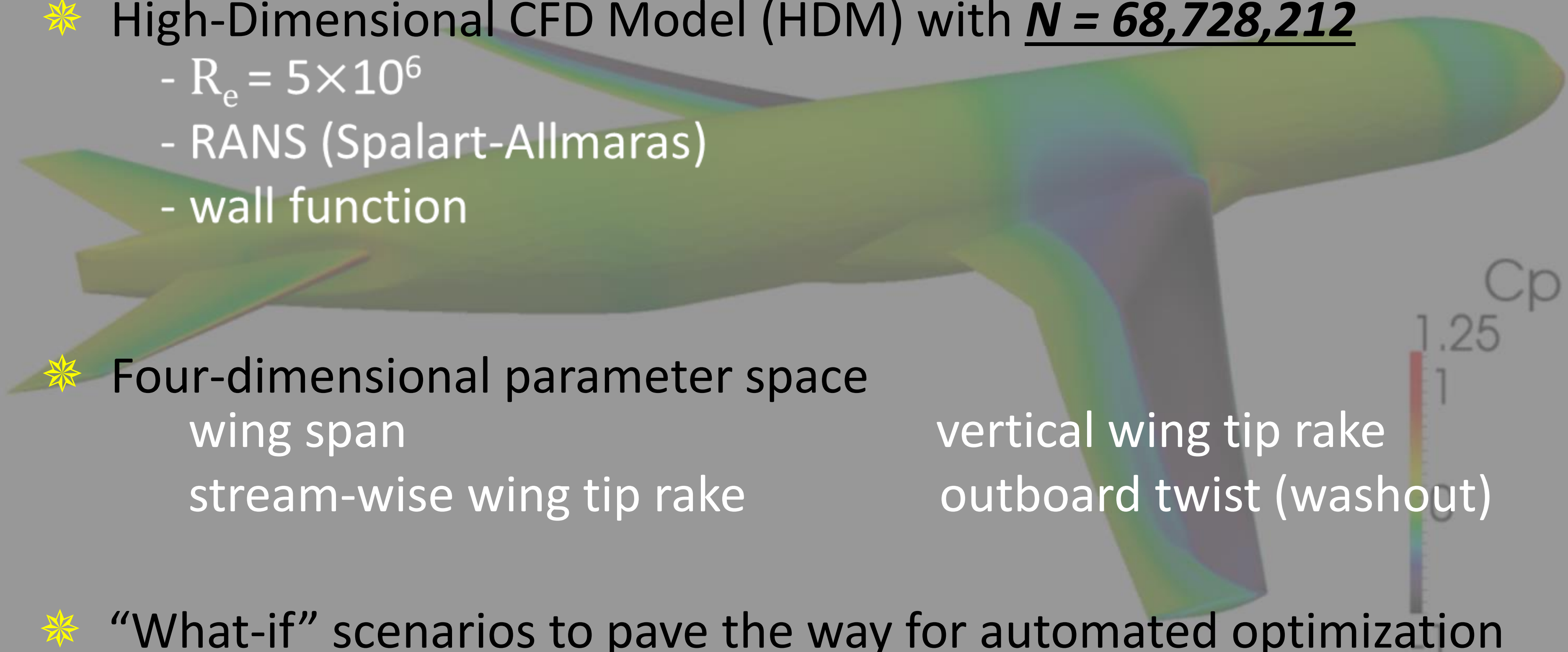
WHAT-IF ANALYSIS

- Cruise conditions

  $M_\infty = 0.85$ angle of attack = 2.32°

  λ-shock

- High-Dimensional CFD Model (HDM) with ***N = 68,728,212***
  - $R_e = 5\times10^6$
  - RANS (Spalart-Allmaras)
  - wall function



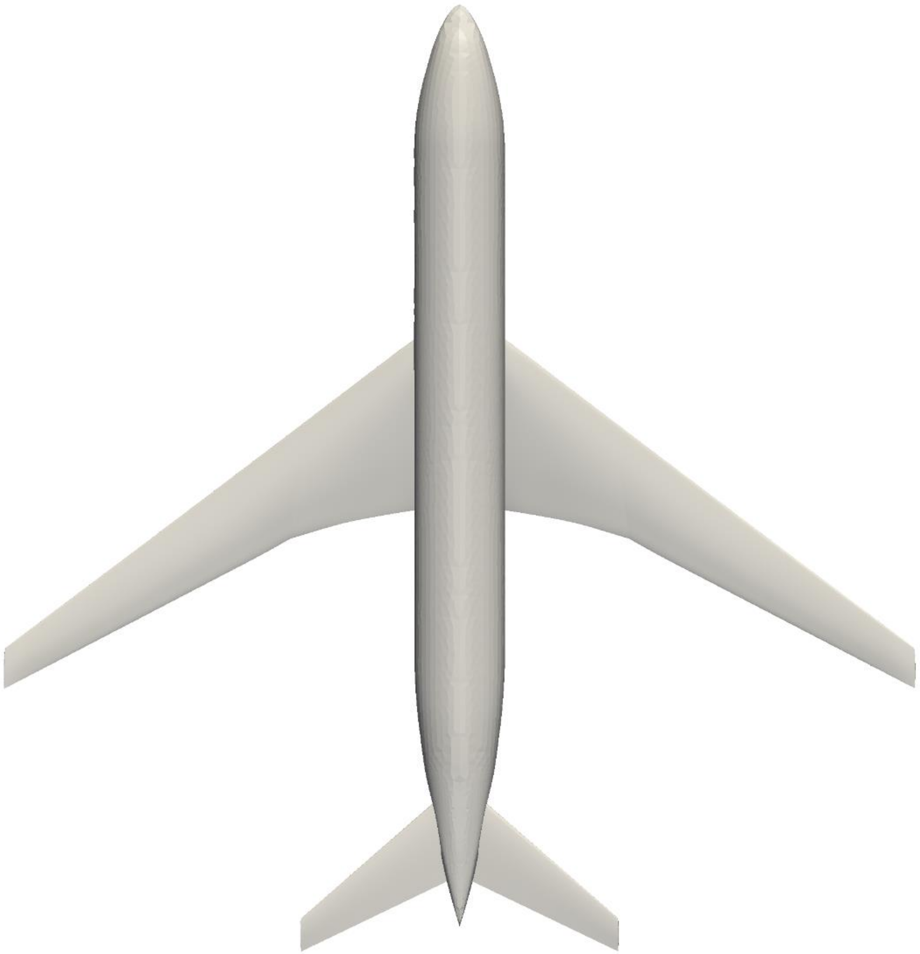
- Four-dimensional parameter space

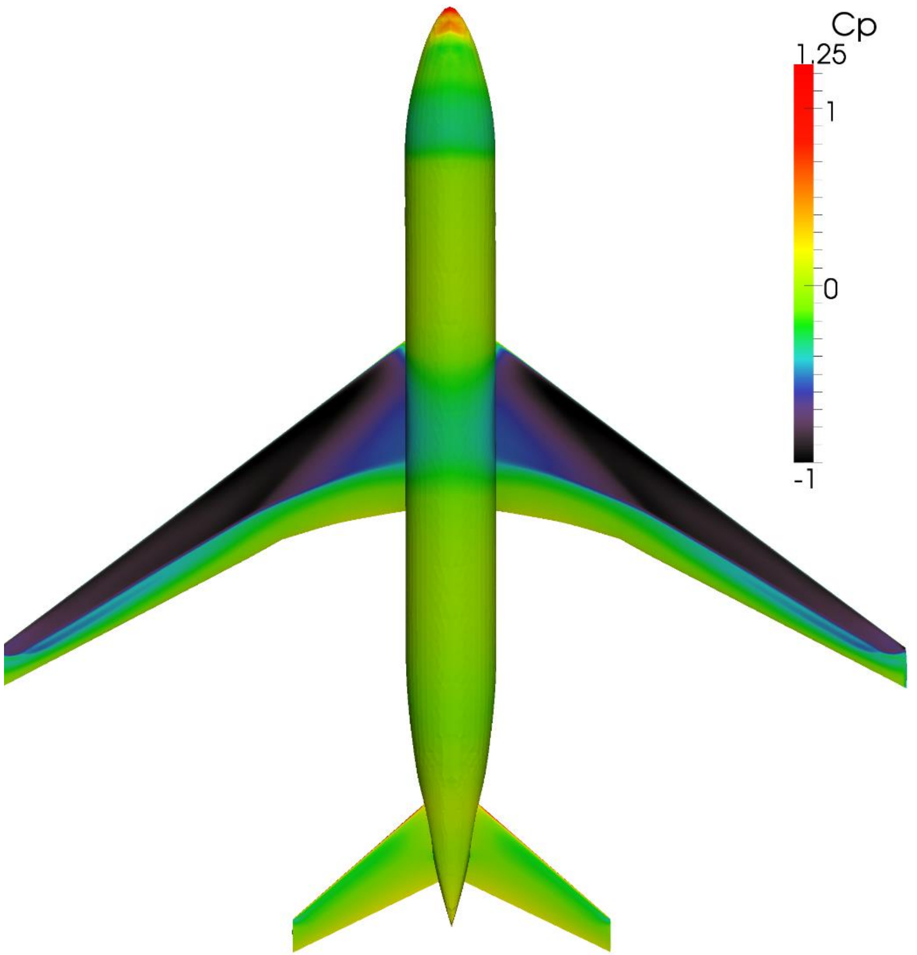
  wing span vertical wing tip rake

  stream-wise wing tip rake outboard twist (washout)

- "What-if" scenarios to pave the way for automated optimization

WING SPAN RANGE



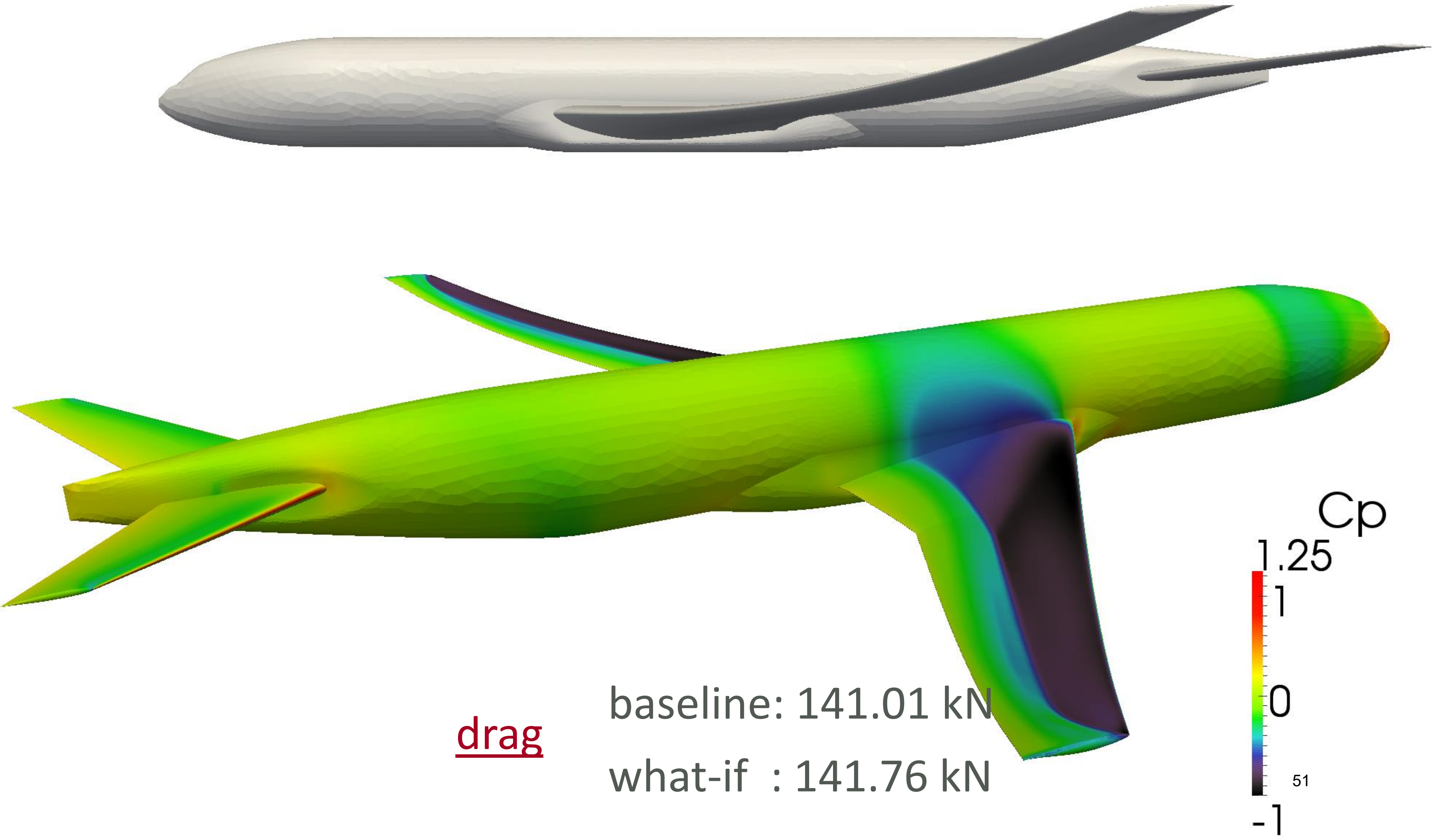


drag

baseline: 141.01 kN

what-if : 151.27 kN

VERTICAL TIP RAKE RANGE

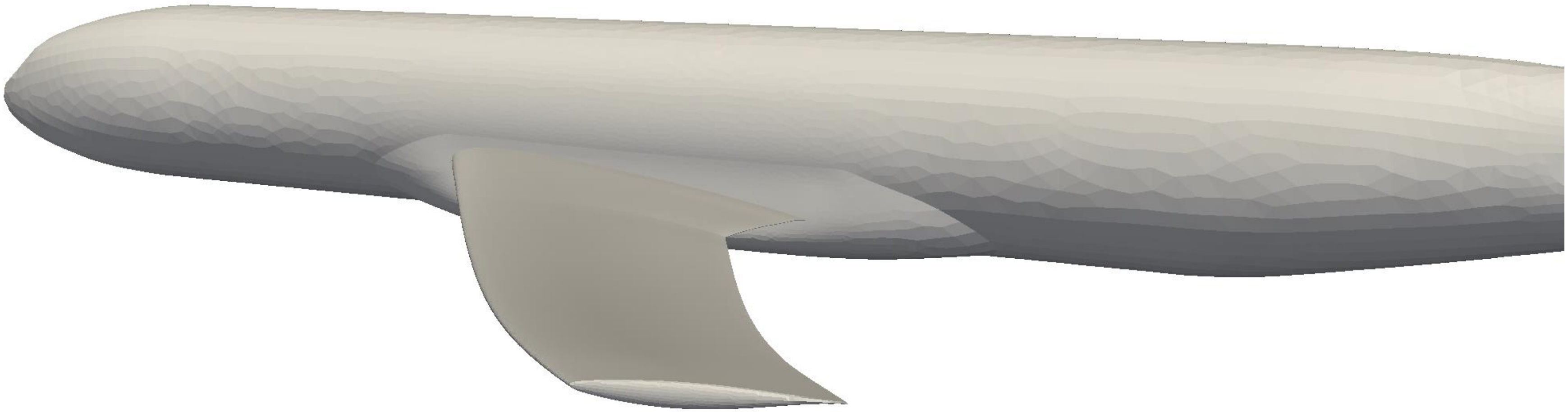


drag

baseline: 141.01 kN

what-if : 141.76 kN

STREAM-WISE TIP RAKE RANGE

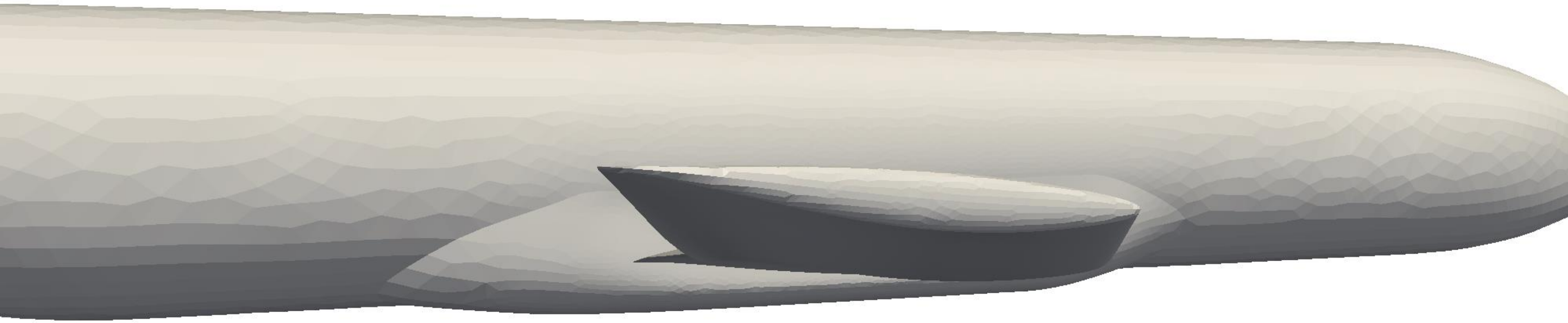


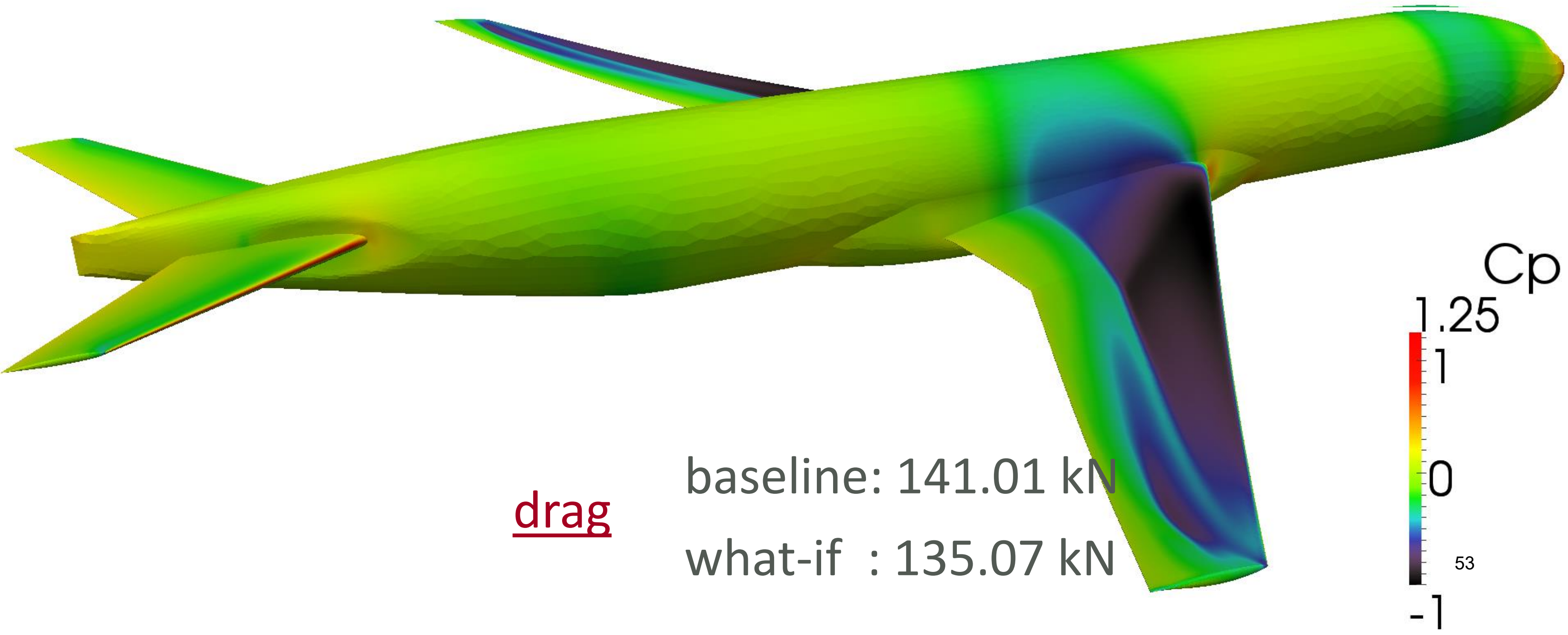
drag

baseline: 141.01 kN

what-if : 139.21 kN

OUTBOARD TWIST RANGE





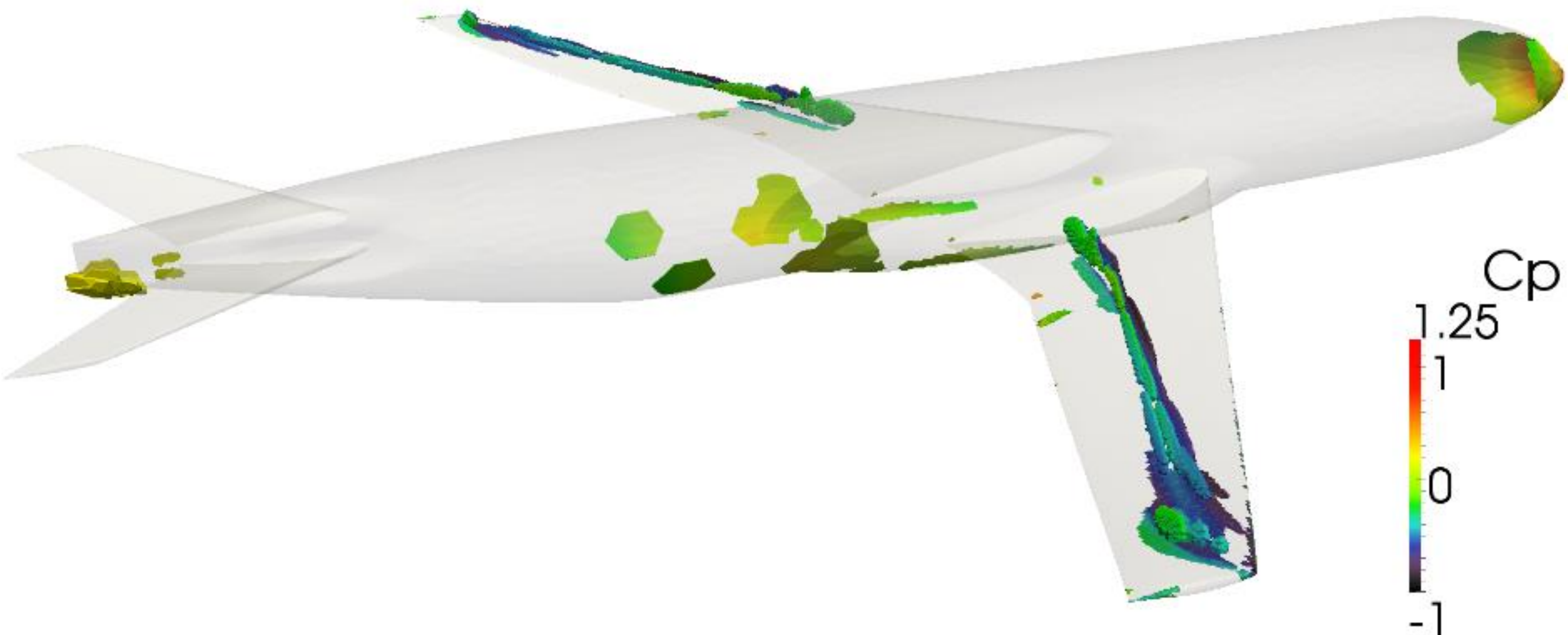
drag

baseline: 141.01 kN

what-if : 135.07 kN

SAMPLING/TRAINING AND GLOBAL PROM CONSTRUCTION

- Sampling at the corners & face centers of the 4D parameter space
  - 24 sampled configurations (1 snapshot / configuration)
  - global ROB of dimension n = 23 (instead of N = 68,728,212)
  - hyper reduction*: 5,000 vertices (instead of 11,454,702)



hyper reduction: reduced mesh computed using a ML approach

COMPUTATIONAL OVERHEAD (OFFLINE)

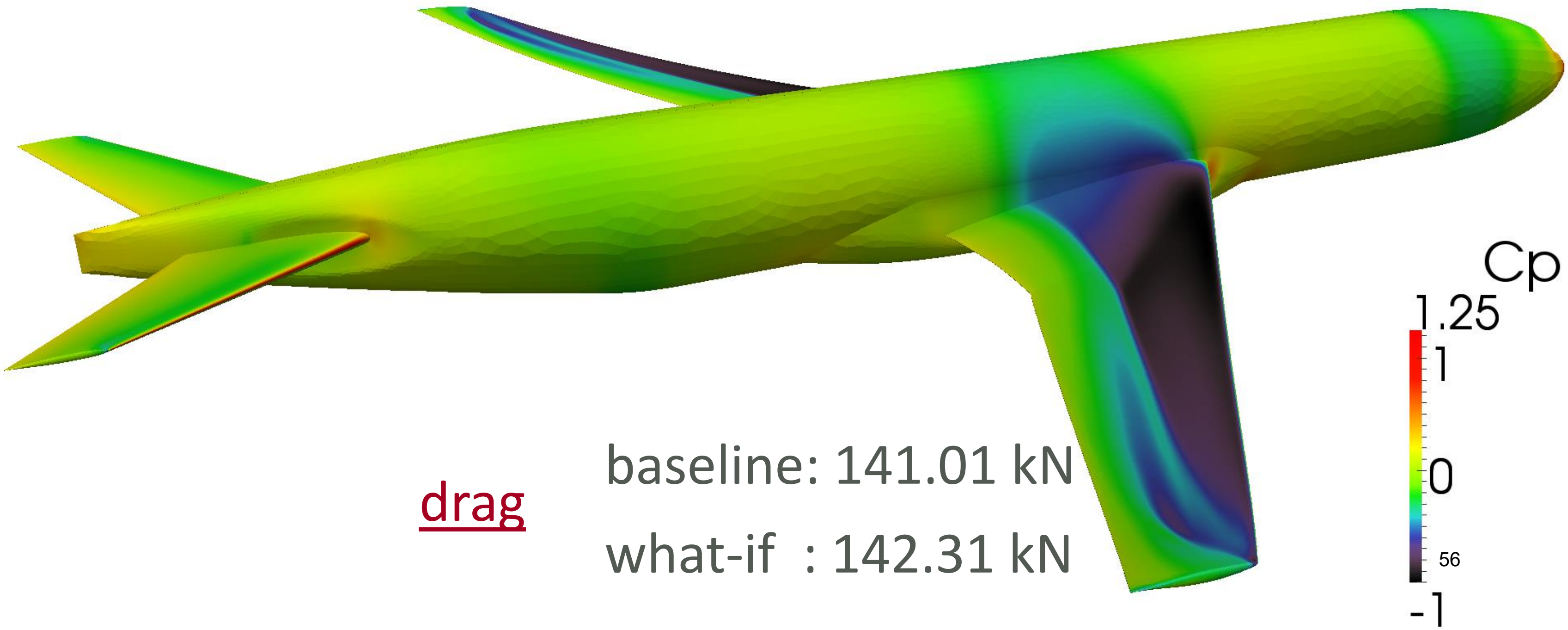
- Excalibur (Cray XC40, ARL)
  - 1,024 cores assigned to each sampled configuration
  - 2 hrs wall-clock time per sampled configuration
  - 11.6 mns wall-clock time for constructing global ROB
  - 3 mns wall-clock time for constructing and hyper reducing the global PROM on 1,024 cores

U.S.ARMY

→ wall-clock time investment: ***2.25 hrs on 24,576 cores***

WHAT-IF ANALYSIS (CASE STUDY#1)

- Example
  - parameter point at the center of the database
  - "real-time" prediction (***laptop***): 29 s (deform reduced mesh)
    30 s (PROM soln)
    78 s (HDM mesh morphing)
    32 s (HDM soln reconstruction)
    170 s (< 3 mins) wall-clock time



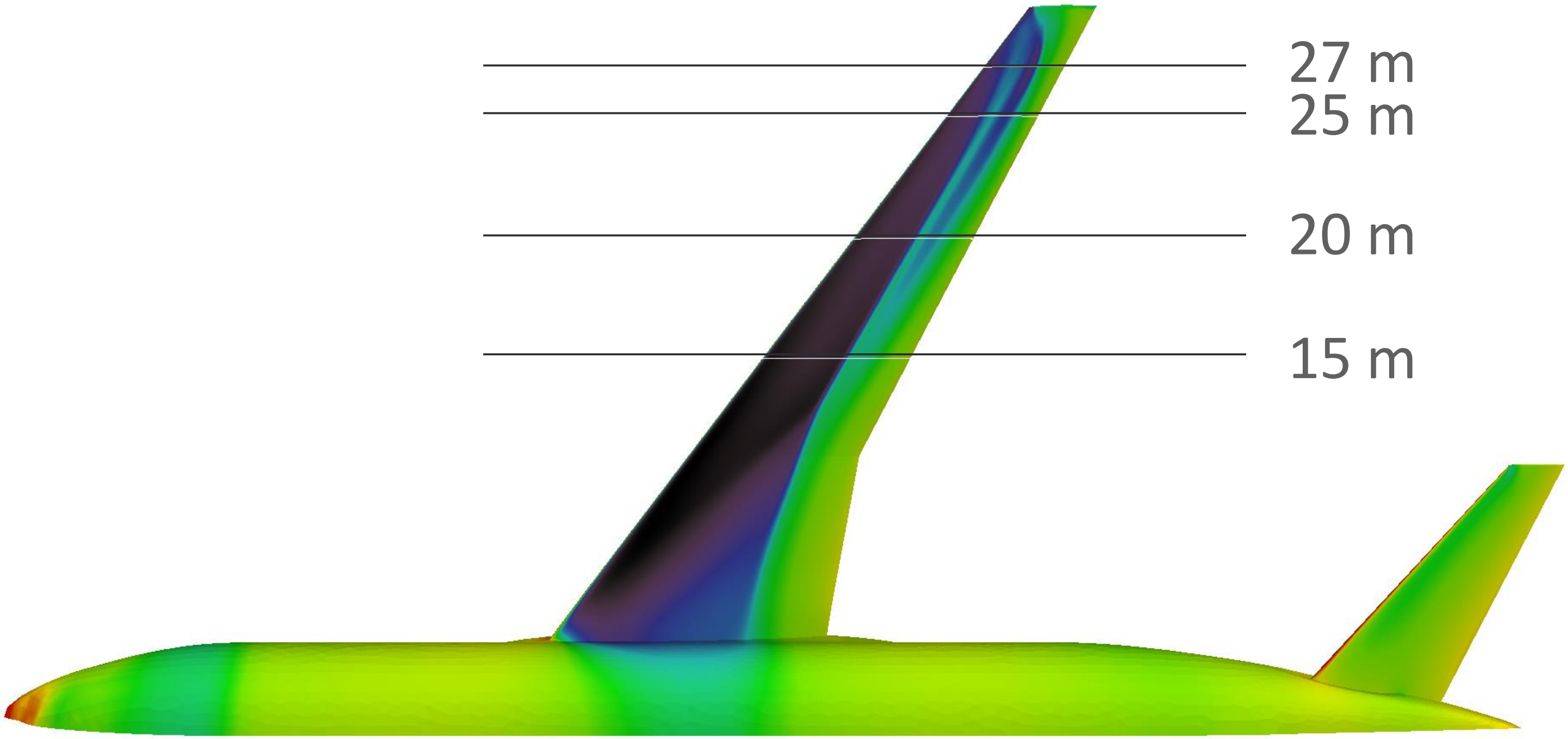
drag

baseline: 141.01 kN

what-if : 142.31 kN

WHAT-IF ANALYSIS (CASE STUDY#1)

- Accuracy of global PROM at the unsampled center of the database
  - C_p contour locations from symmetry plane



WHAT-IF ANALYSIS

- Accuracy of global PROM at the unsampled center of the database