# On the potential for high throughput high fidelity simulations

Sanjeeb T Bose

CASCADE TECHNOLOGIES

F. Ham, C. Ivey, M. Emory, L. Shunn,
G. Bres, K. Wang, D. Philips, D. Kim,
P. Quang



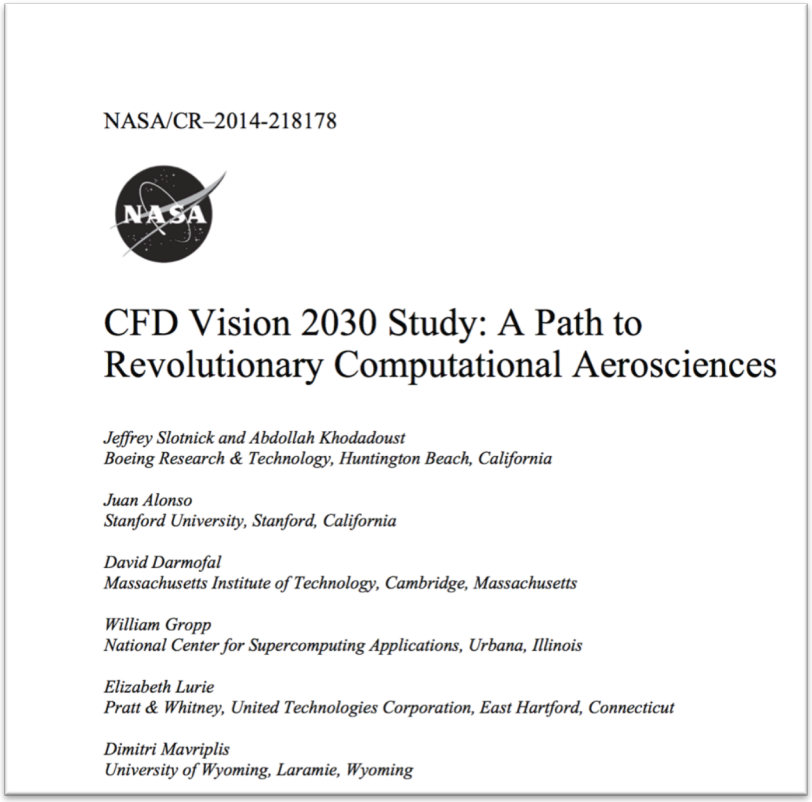

K. Goc, P. Moin

Acknowledging support from:

Argonne NATIONAL LABORATORY
OAK RIDGE National Laboratory
NAVAIR
NASA
U.S. DEPARTMENT OF ENERGY
BOEING

Advanced Modeling & Simulation (AMS) Seminar Series
NASA Ames Research Center, June 17th,2021

# Promise & obstacles of (WM)LES for external aerodynamics
## NASA Vision 2030 (2014)



NASA/CR–2014-218178

NASA

CFD Vision 2030 Study: A Path to Revolutionary Computational Aerosciences

*Jeffrey Slotnick and Abdollah Khodadoust*
*Boeing Research & Technology, Huntington Beach, California*

*Juan Alonso*
*Stanford University, Stanford, California*

*David Darmofal*
*Massachusetts Institute of Technology, Cambridge, Massachusetts*

*William Gropp*
*National Center for Supercomputing Applications, Urbana, Illinois*

*Elizabeth Lurie*
*Pratt & Whitney, United Technologies Corporation, East Hartford, Connecticut*

*Dimitri Mavriplis*
*University of Wyoming, Laramie, Wyoming*

Full flight envelopes encounter flow phenomena that remain challenging to predict (e.g., BL separation)

Other applications benefit from fundamentally unsteady methods (e.g., aeroacoustics)

Utilization of LES in design settings impeded by computational cost, algorithmic robustness in complex geometries, unknown predictive capabilities in complex flow

Overcoming obstacles necessary for LES to complement multi-fidelity simulations in design environment

# Pacing questions in light of NASA Vision 2030

WMLES in this Re range *might* be possible with full utilization of leadership (petaflop/exaflop) systems in O(day)

V2030 grid point estimates for WMLES of wing geometry

| $Re_c$ | $N_{grid}$ (AR = 1) | $N_{grid}$ (AR = 3) |
|---|---|---|
| $10^6$ | 9B | 27B |
| $10^7$ | 85B | 255B |

Resolution requirements are not sensitized to quantity of interest

Is 10-100X increase due to laminar BL resolution necessary?

What is trade-off between resolution and error? Is it purely viscous (Re) scaled?

What are the impacts on costs due to advent of accelerated architectures (e.g., GPUs) on time/$?

# Outline

- Resolution sensitivities in free shear flows
- High lift predictions from JAXA Standard Model
- Compressor stall limits for NASA Rotor 37

Resolution requirements are not sensitized to quantity of interest

Is 10-100X increase due to laminar BL resolution necessary?

What is trade-off between resolution and error? Is it purely viscous (Re) scaled?

- Time to solution on GPU-based clusters
- Cost ($) impact of accelerated architectures

What are the impacts on costs due to advent of accelerated architectures (e.g., GPUs) on time/$?

# Simulation environment (charLES) & caveats

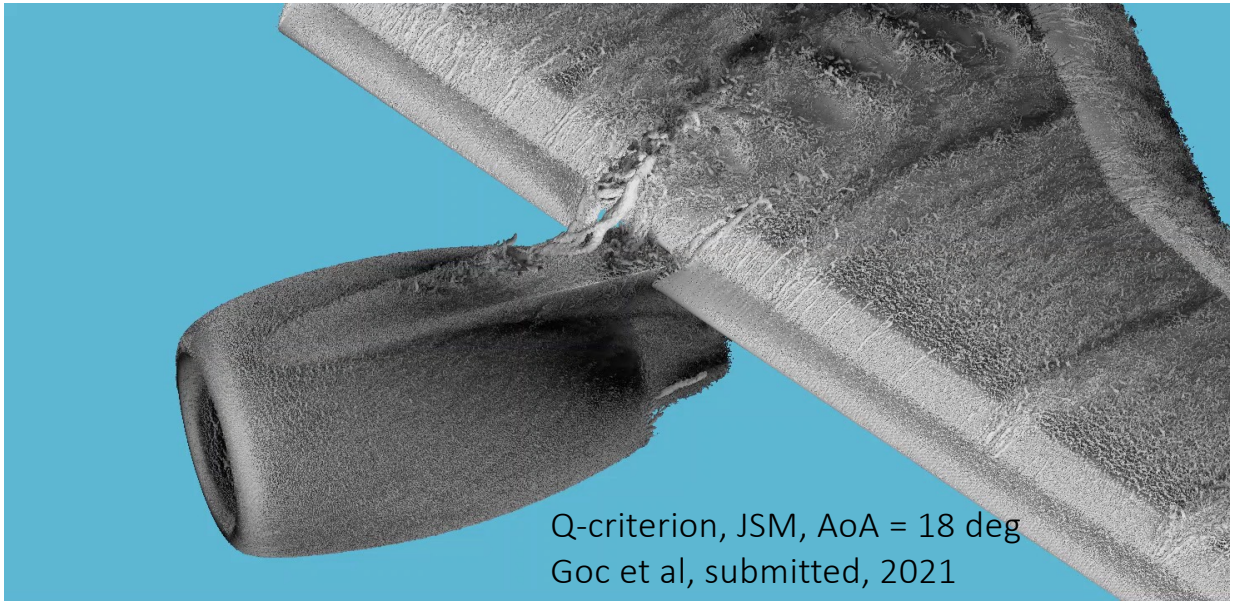
All calculations presented are performed in charLES

Grids generated based on 3D clipped Voronoi diagrams

Second-order, unstructured, approximately entropy preserving spatial discretization (Bres et al 2018)

Explicit RK SSP time advancement (Gottlieb et al 2001)

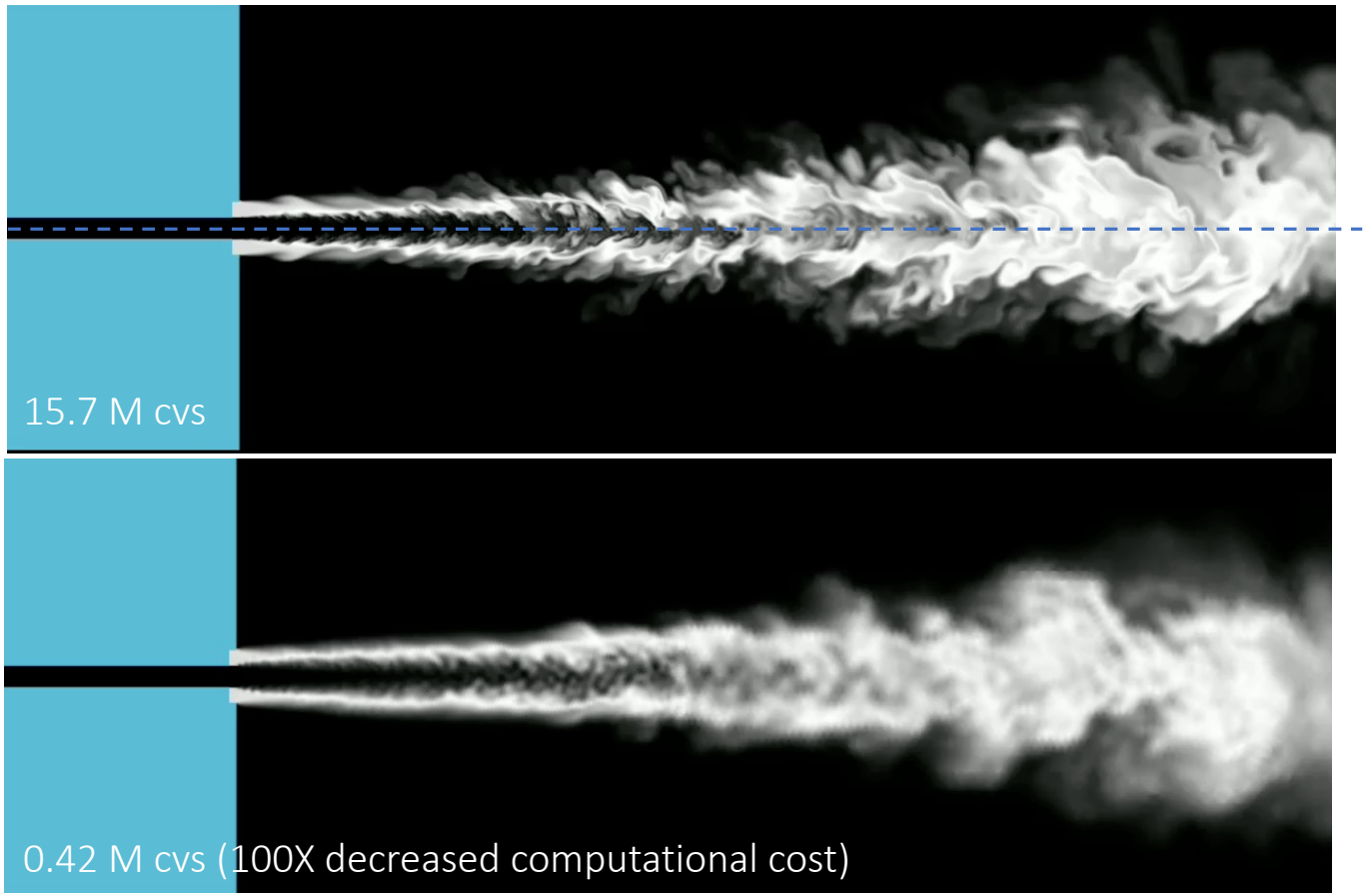
Equilibrium wall models (Bose & Park 2018; Larsson et al 2016); Vreman SGS (Vreman 2004), Dyn. Smagorinsky (Germano et al 1991)




All conclusions (resolution/GPU acceleration) based on charLES

# Resolution requirements for free shear flows

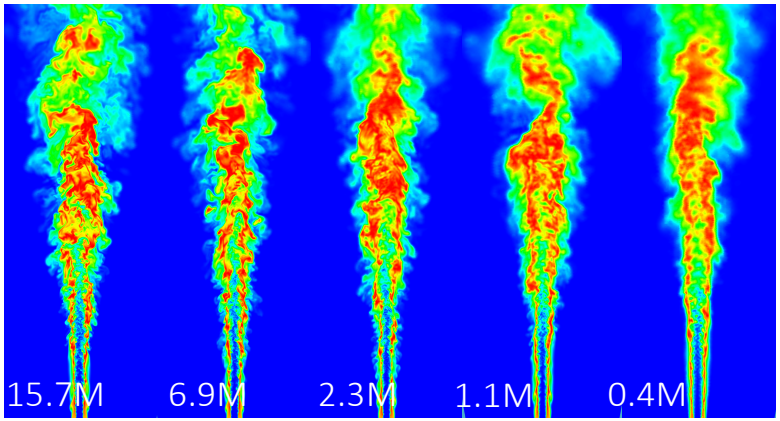
## Sandia flame D (Barlow et al, 2004-2005)

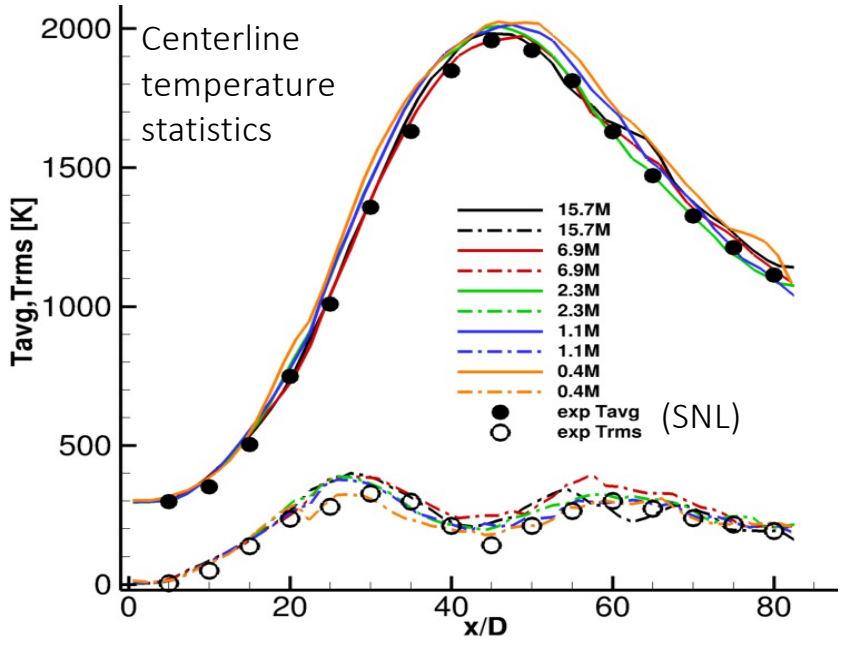



- $Re_D \approx 20{,}000$, piloted/methane air flame
- Use a reacting flow case in order to stress nonlinear stability of numerical solver in simple free shear flow
- Predominant quantity of interest: first/second order statistics of temperature field

# Centerline temperature mean/rms show collapse across multiple resolutions

Instantaneous temperature fields

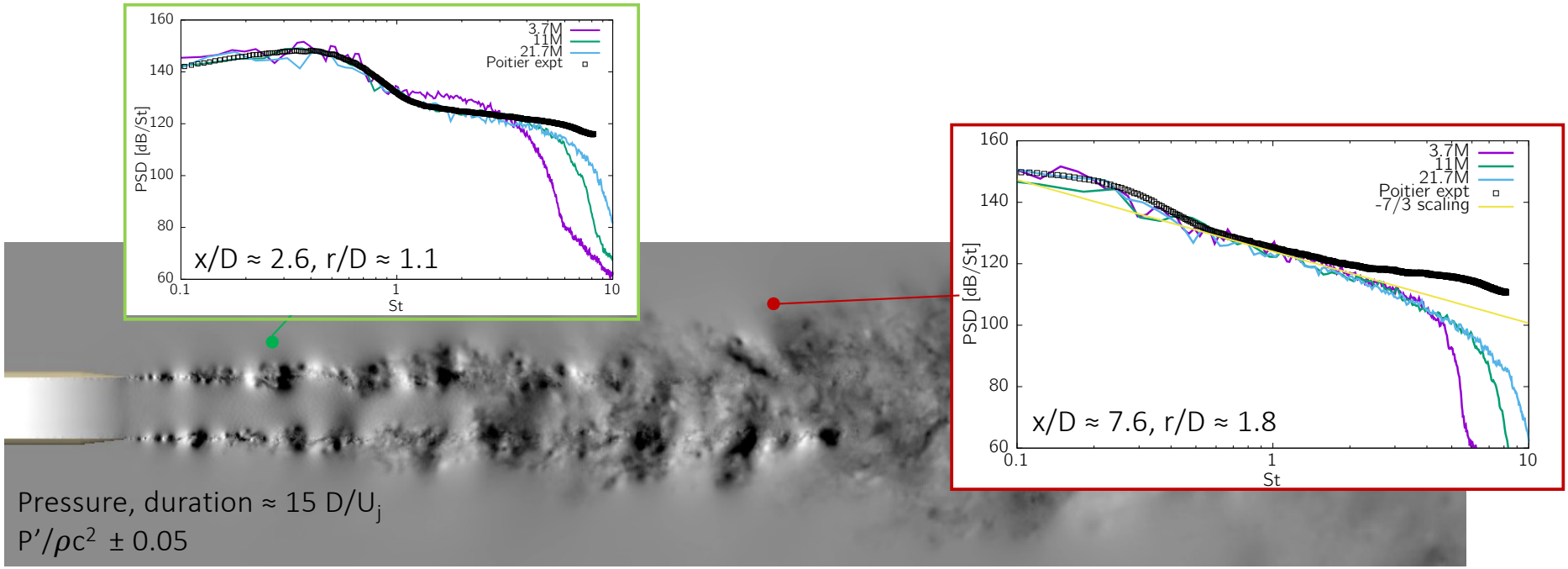



- All resolutions considered here show relatively robust avg/rms statistics (will predict similar flame length)
- Further reduction of grid resolution will violate either large scale resolution limits (initial shear layer/BL thickness or integral length scales of the jet)

# Convergence of turbulent broadband spectra in free shear flows

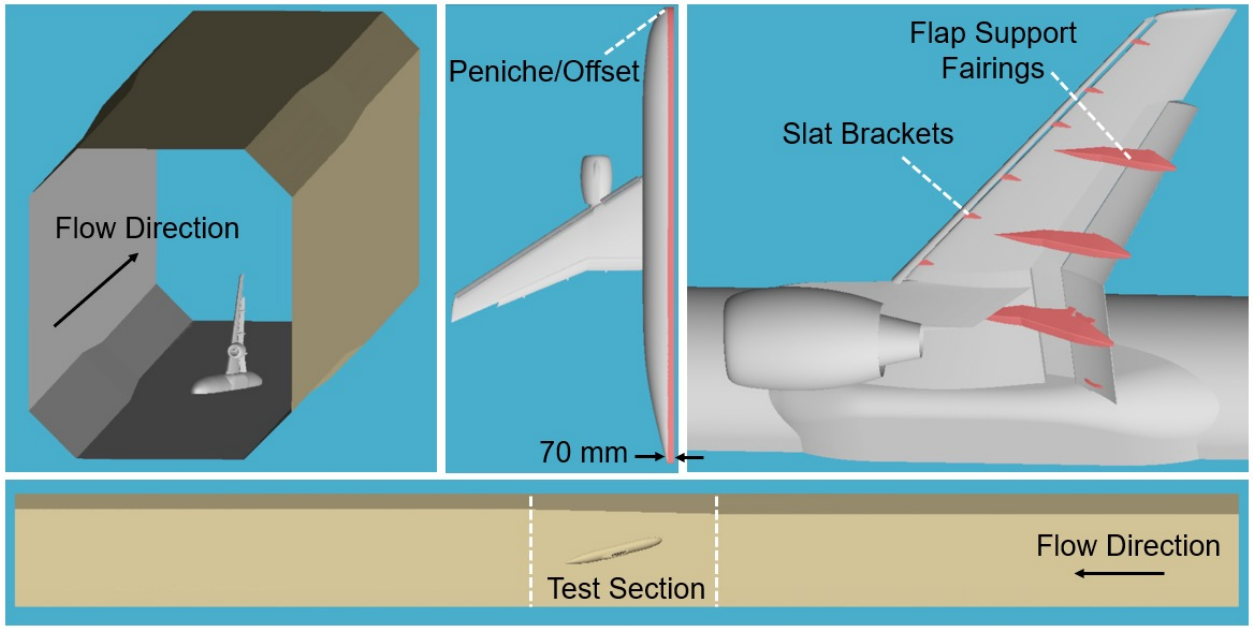
## Near field pressure spectra, M = 0.9, $Re_j = 10^6$ (Bres et al, JFM 2018)




Overall pressure spectra collapse for resolutions with marginal shear layer resolution; broadband roll-off consistent with grid Nyquist limits

# Predictions of high Reynolds number wall bounded flows

## JAXA Standard Model, $Re_c$ = 2M (Yokokawa et al 2009)



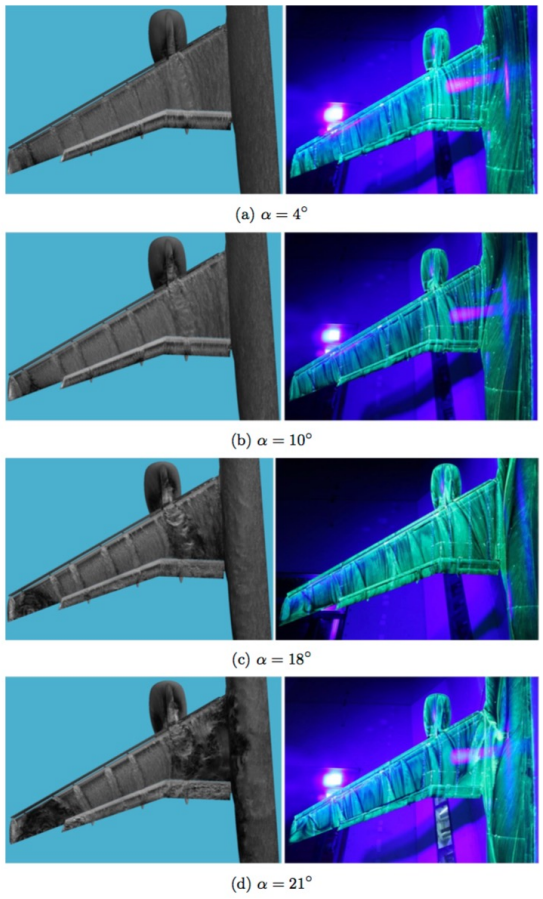

Goc et al, submitted, 2021

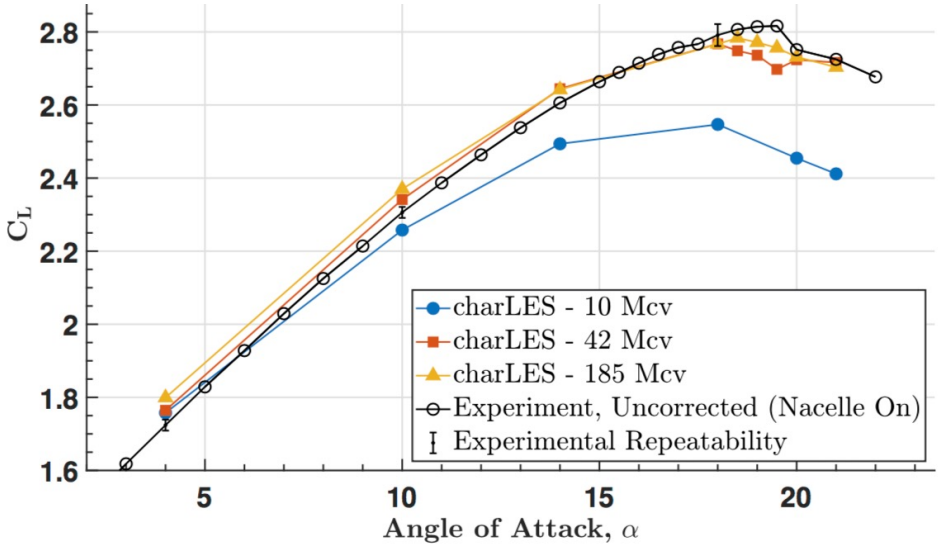
Simulations of WBNP configuration in the wind tunnel; comparison against uncorrected experimental data

Primary quantities of interest are integrated forces (lift/drag/moment)

Grid sequence considered from $\delta/\Delta \approx$ 2-5 (coarsest) to 10-20 (finest) based on attached flow on main element

# Lift polar comparison as a function of resolution



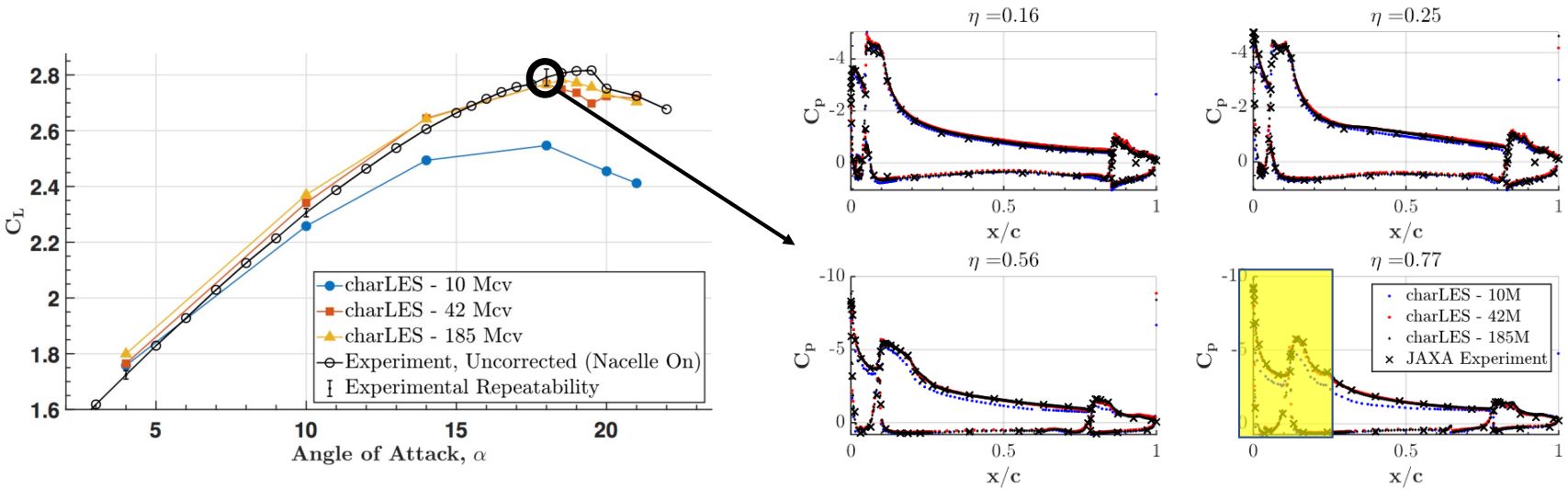




Med/fine resolution show near convergence in lift coefficient, with predictions nearing experimental repeatability bounds

Absolute predictions approaching accuracy targets (Clark et al 2020)

Grid estimate (V2030): O(60B) cells; present simulations *$10^2$-$10^3$X* coarser than estimates

# Where is the source of resolution sensitivities for the lift?




- Two reasons to believe that the convergence is not just viscous resolution related [≠ f(Re) alone]:
  - Resolution sensitivity in lift is not uniform across all angles of attack – more severe near $C_{lmax}$
  - $C_p$ sensitivity most pronounced at outboard wing stations experiencing strongest acceleration

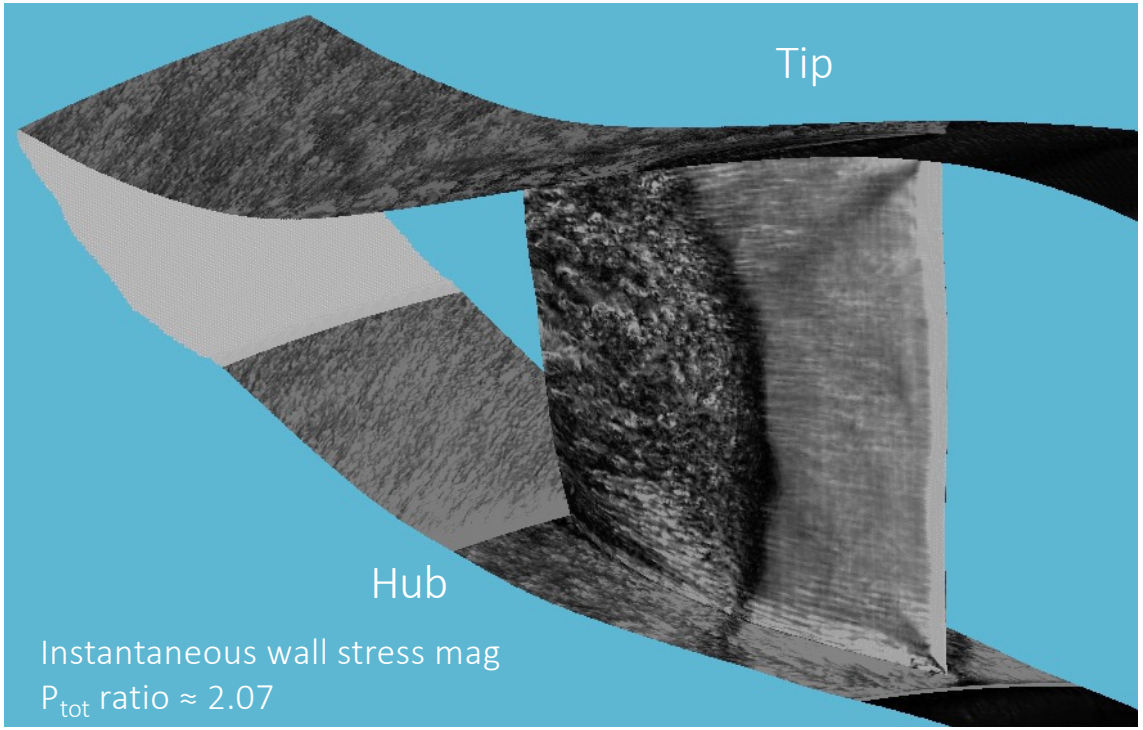
# Application to transonic and transitional flows

## 100% design speed, NASA Rotor 37 (Suder 1997)

Benchmark validation case for a realistic axial compressor stage

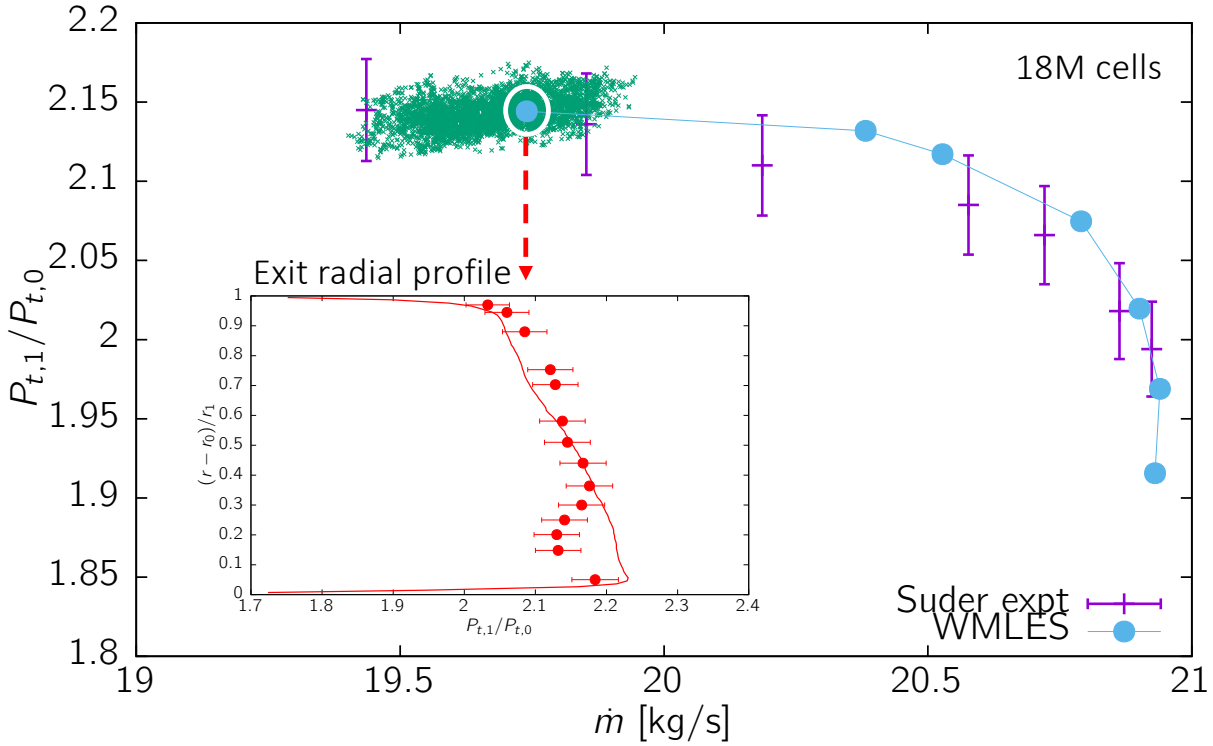
$M_{rel}$ = 1.1 and 1.5 at hub and tip,
$Re_c \approx O(1.5 \times 10^6)$

Objective to characterize performance map (total pressure vs mass flow rate, from choke to stall)



# NASA Rotor 37 speedline predictions

## 100% design speed comparison against Suder expt (1997)




LES captures performance characteristics of compressor map (including approach of the stall boundary)

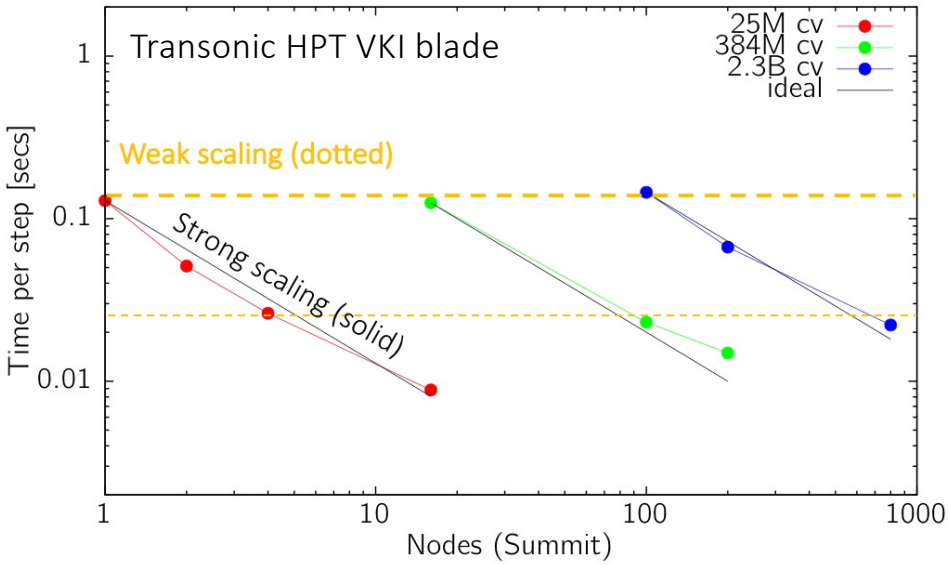
Grid point reduction of similar magnitude compared against V2030 estimates

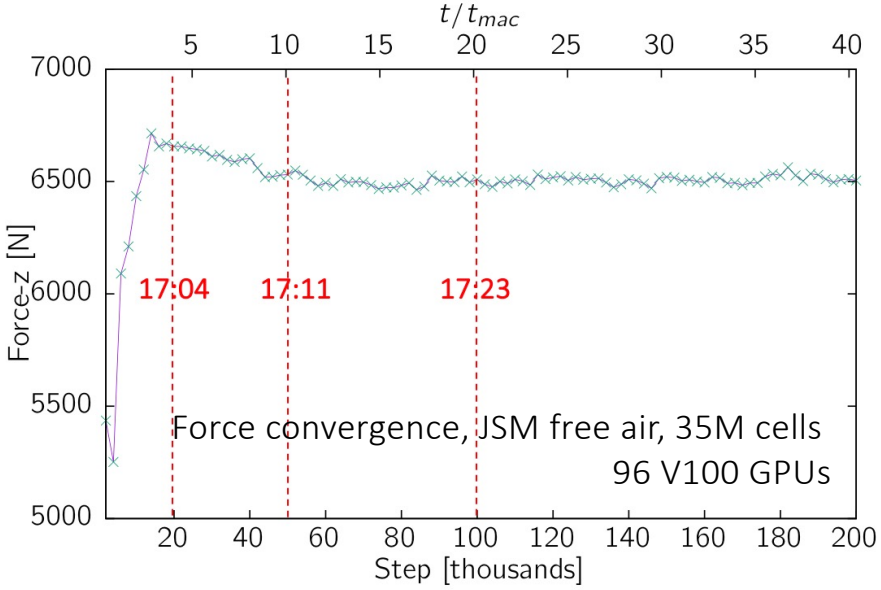
# Remarks about the computational cost envelope

- LES calculations herein have room for improvement (meeting accuracy requirements for lift/drag, or prediction of peak adiabatic efficiency), but…
- Cases considered show encouraging agreement with experiment for critical quantities of interest with > 10X reduction compared to V2030 estimates
- Wall bounded cases considered at medium resolutions (20M - 50M cells) consume 20-60K core-hours; turn-around times possible within a day on modest computing resources
- *What is the potential reduction in computational cost from compute architecture enhancements?*

# Time to solution on massively parallel GPU architectures

## strong/weak scalability on ORNL Summit

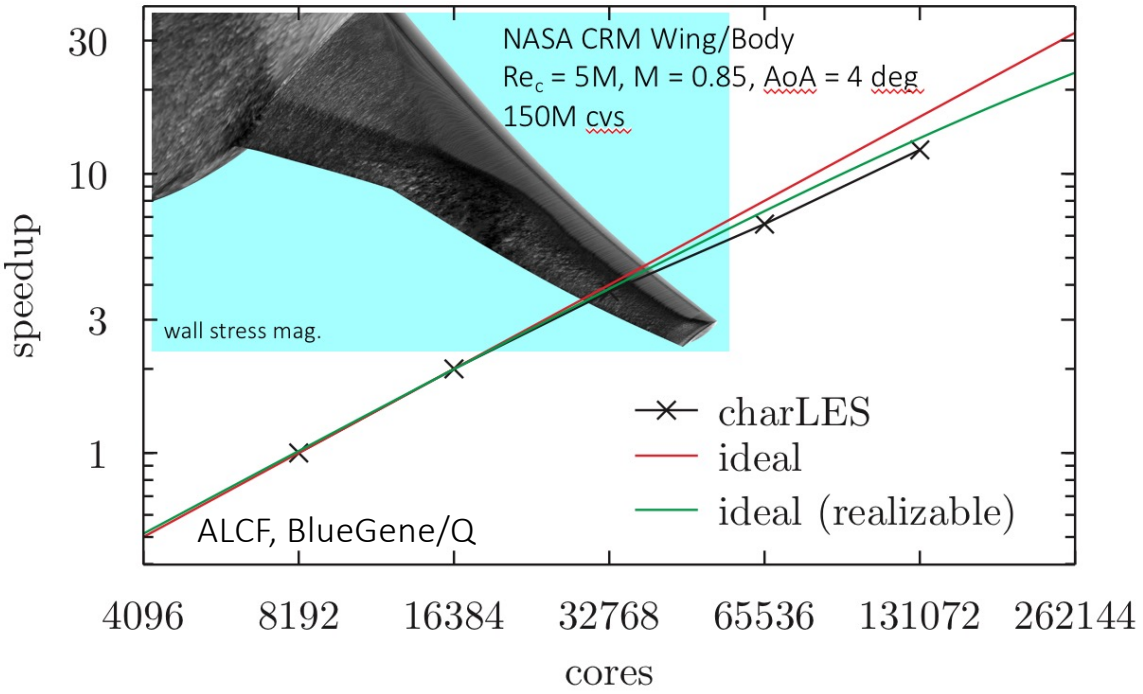



- charLES solver ported to run on AMD/NVIDIA GPUs; good strong parallel efficiency up to O(B) cells (> 80%)
- At the strong scalability limit, cases considered thus far can complete in *O(minutes)-O(hours)*

# Is GPU acceleration faster than traditional CPU architectures?

## Consideration of resource constraints



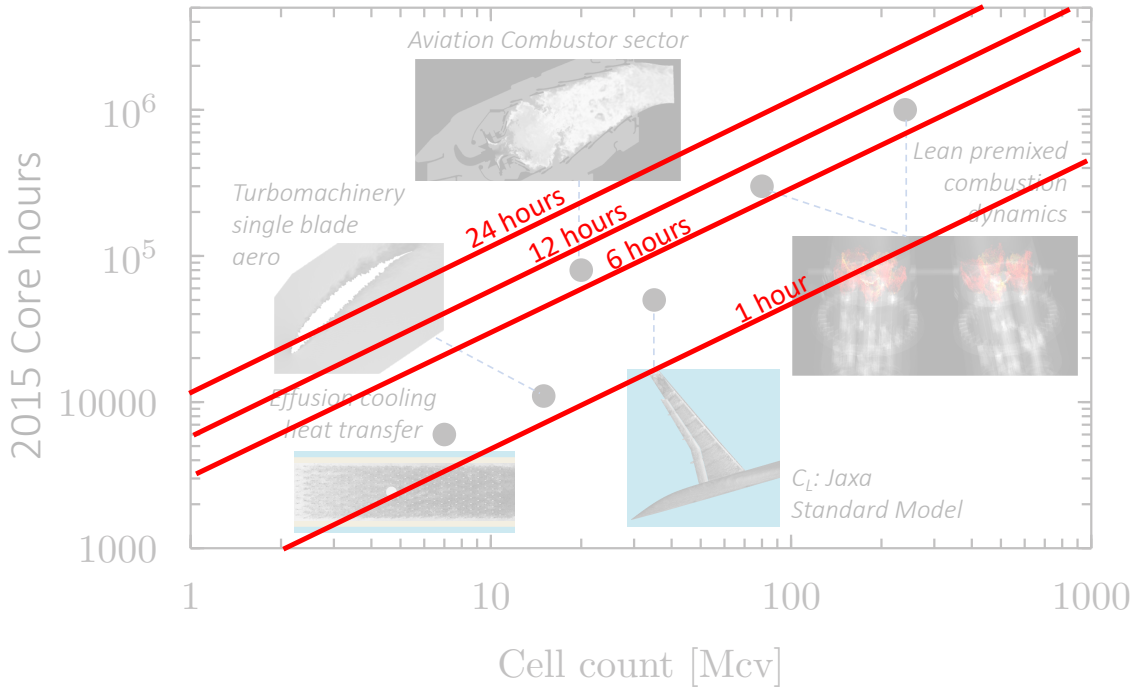

Prevailing narrative is that GPUs are fast compared to CPU (inappropriate comparison)

Parallel efficiency for CPU solver persist to loads of O(1000) cells/rank

At CPU scalability limit, solutions also complete in O(hours)

Why don't we compute like this now? Usually resource constrained…

# Time to solution at the scalability limits (II)

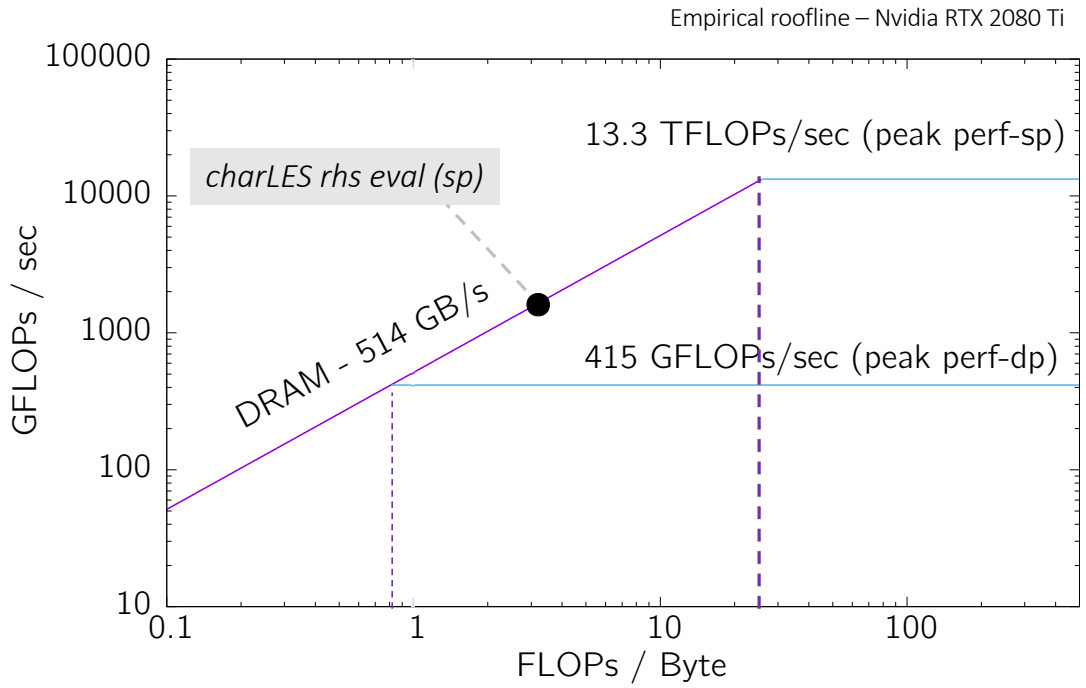



Variety of cases computed at the scalability limits could be computed within 1 day

Would require running 50M cell grid on 25K cores – not likely in industrial settings

Better question: do GPUs ameliorate resource constraints because they are cheaper ($)?

# Pricing hardware efficiency ($) based on algorithmic constraints
## using roofline analysis (Williams et al 2009)



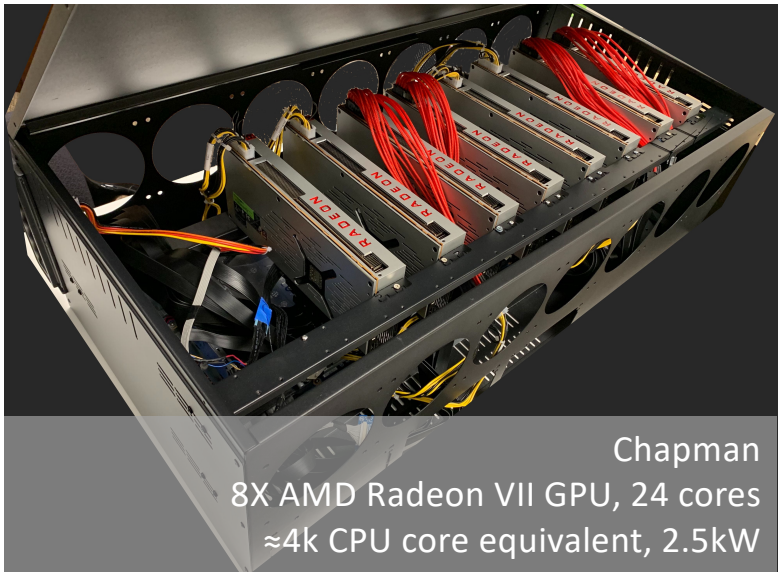

Throughput for charLES calculations limited by available memory bandwidth; leading order cost efficiency from mem. bw/$

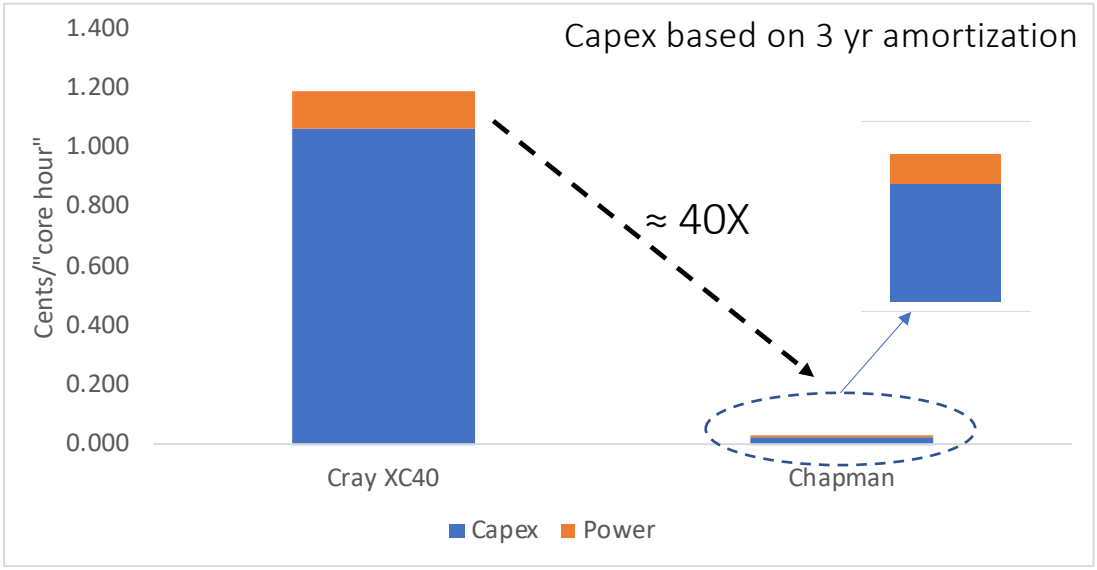
Achieving this is complicated in practice – on GPU, consideration of mixed precision, aggressive latency hiding, etc…

But, how cheaply could we compute if attempt to $-cost optimize the hardware?

# Estimating the floor for cost ”optimal” hardware

## based on equivalent throughput for charLES



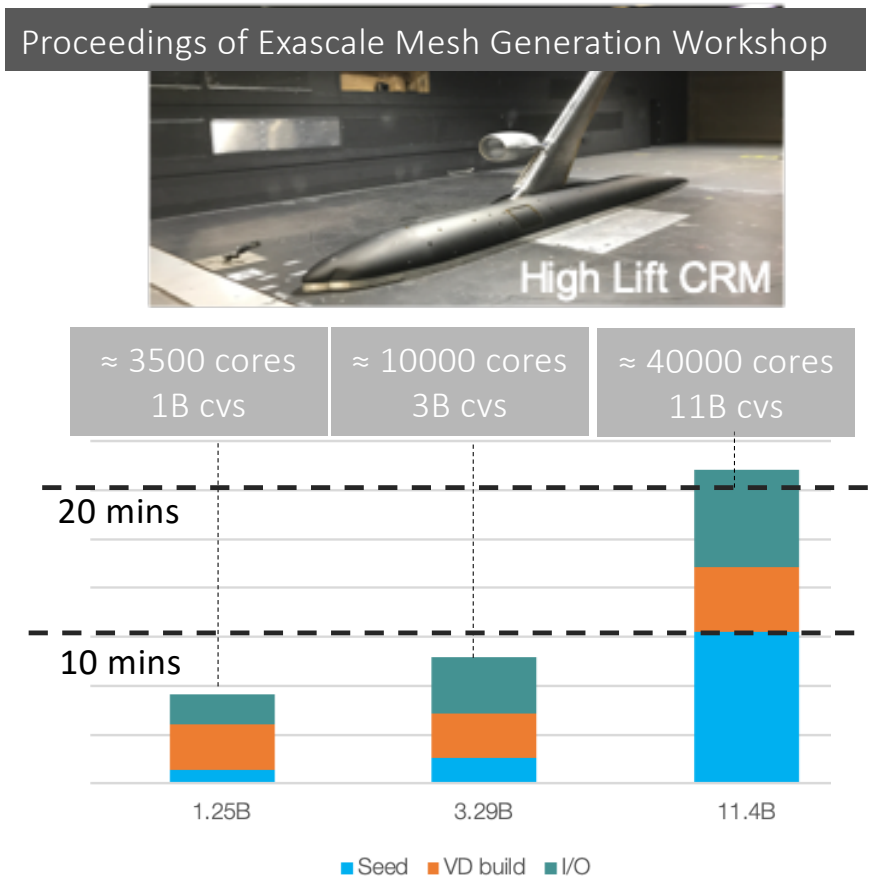




Cost reductions at equivalent throughput achieved ≈ 40X; gains evenly split between GPU architecture suitability (e.g., IPC) and optimizing memory bw/$

Assessing power consumption is becoming increasingly important with capex cost reductions

# Addressing other bottlenecks for time to solution
## Scaling the pre-processing workflow (fast mesh generation)

- If solutions complete in O(hours), the rest of the workflow must also accelerate accordingly
- Feasible to generate sufficiently large grids on complex geometries to accommodate (estimated) resolution for flows at realistic Reynolds numbers [$10^7$-$10^{10}$ grid cells]
- Grid generator scales to building grids on O($10^5$) cores in < 30 minutes



# Concluding remarks

- Combination of advances in numerical methods, wall/SGS modeling, and mapping can potentially reduce the computational cost of LES simulations by > 100X

- Resolution estimates presently offered likely incomplete; initial estimates likely too pessimistic
  - Estimates need to be connected to particular quantities of interest
  - Estimates that function of Re alone potentially insufficient

- High throughput environments will require increasing automation (error estimates, automatic grid generation) and robustness