

Методы обработки и анализа больших данных в задачах вычислительной аэрогидродинамики

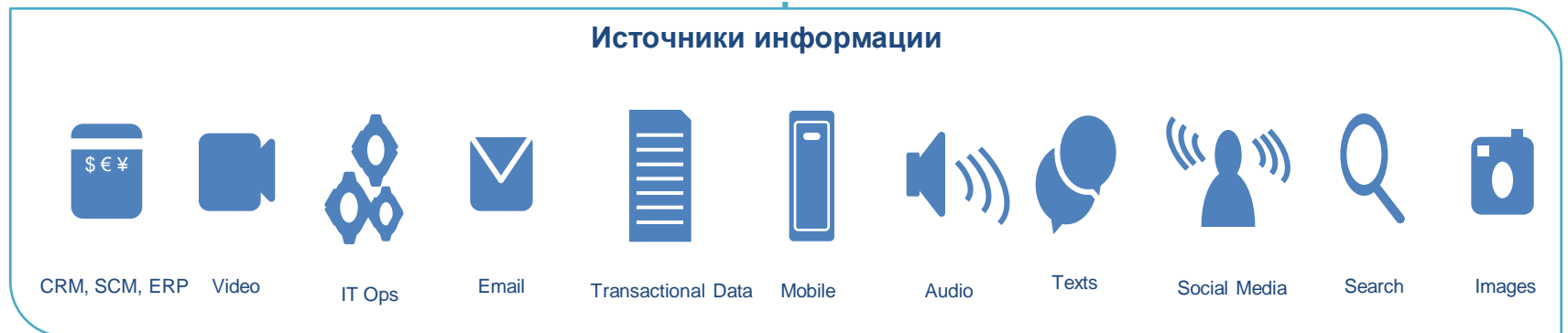
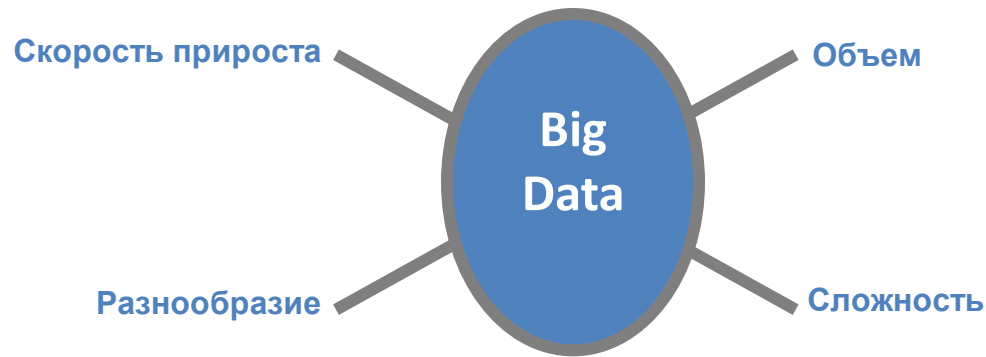
Сергей Стрижак

“Облачные вычисления: образование,
исследования, разработки”

Москва, 5-6 декабря, 2013

Что является Большими Данными (Big Data)?

Данные у которых объем, скорость, многообразие, захват и сложность выходят за рамки возможности использования стандартного ПО для управления, анализа, хранения



Большие данные в науке?

Астрофизика (телескоп “Хаббл”, снимки доступной области неба) ~ 25 TB за 15 лет

Биоинформатика (ДНК-секвенирование генетических последовательностей)

Геофизика и геология (съемка данных с местности)

Данные с метеостанций (температура, скорость)

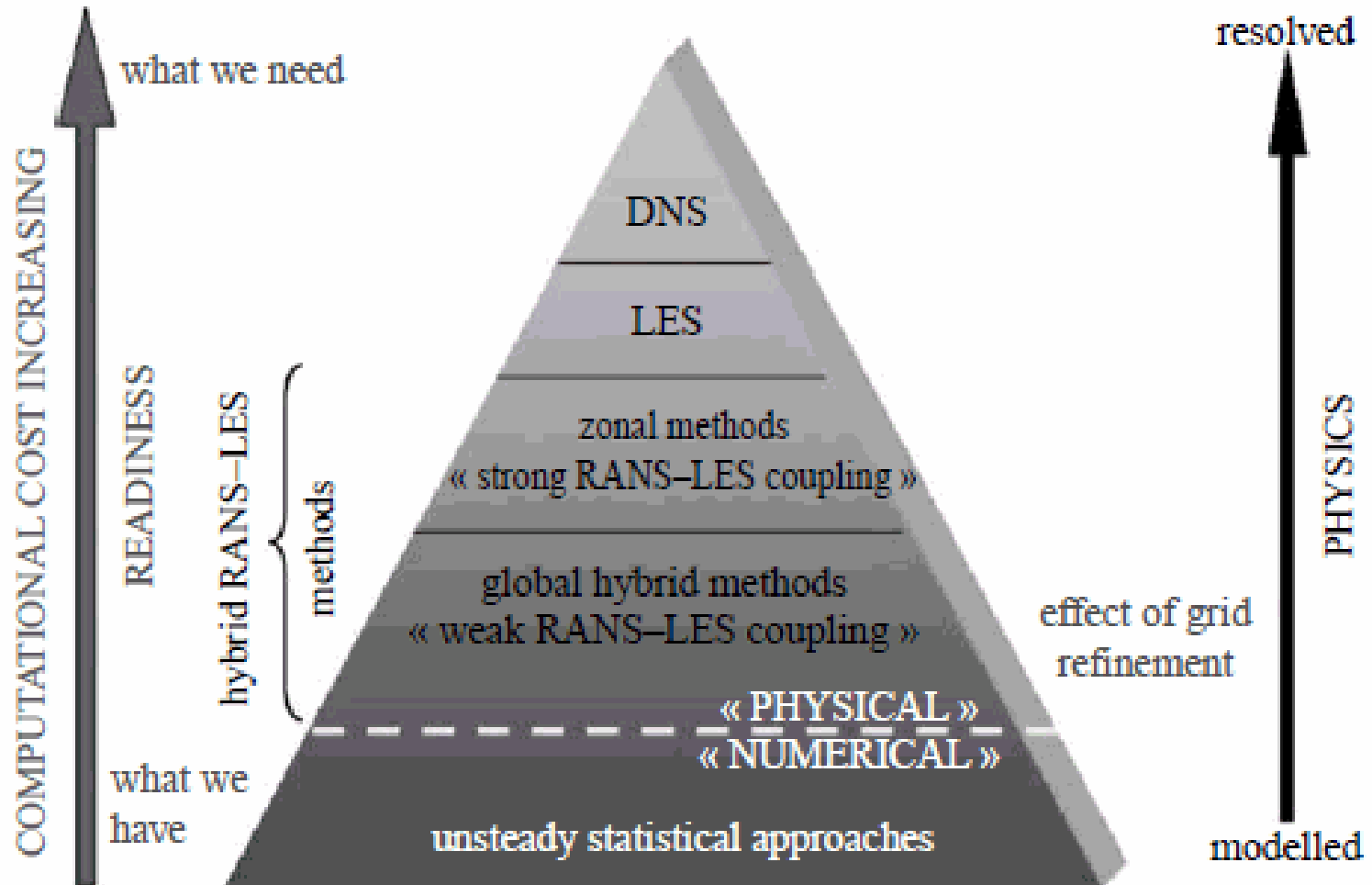
Данные с космических спутников (карты, снимки, ДЗЗ)

Физика элементарных частиц (Большой Адронный коллайдер в CERN, регистрация столкновение ионов тяжелых металлов 4 детекторами) ~ 25 PB в год

Аэрогидродинамика и теплообмен (расчет и эксперимент)

Howe D. et al. Big data: the future of biocuration. Nature. 2008. v. 455.

Различные подходы в вычислительной аэрогидродинамике



Количественные характеристики в вычислительной аэрогидродинамики

N%	Параметр	Значение
1	Интегральные силы и моменты	X, Y, Z, Mx, My, Mz
2	Средние аэродинамические поля	скорость, давление, температура
3	Вторая статистическая характеристика турбулентности	корни квадратные из средних квадратичные пульсации скорости
4	Одноточечный спектральный анализ	спектральная плотность мощности, тензор моментов корреляция,
5	Двухточечный спектральный анализ	когерентность
6	Частотно-временной анализ	время, частота

Задачи оптимизации и неопределенности количественных величин

Проект в рамках web-лаборатории UniHUB. НИР по аэроакустике. Пример “rod-airfoil”.

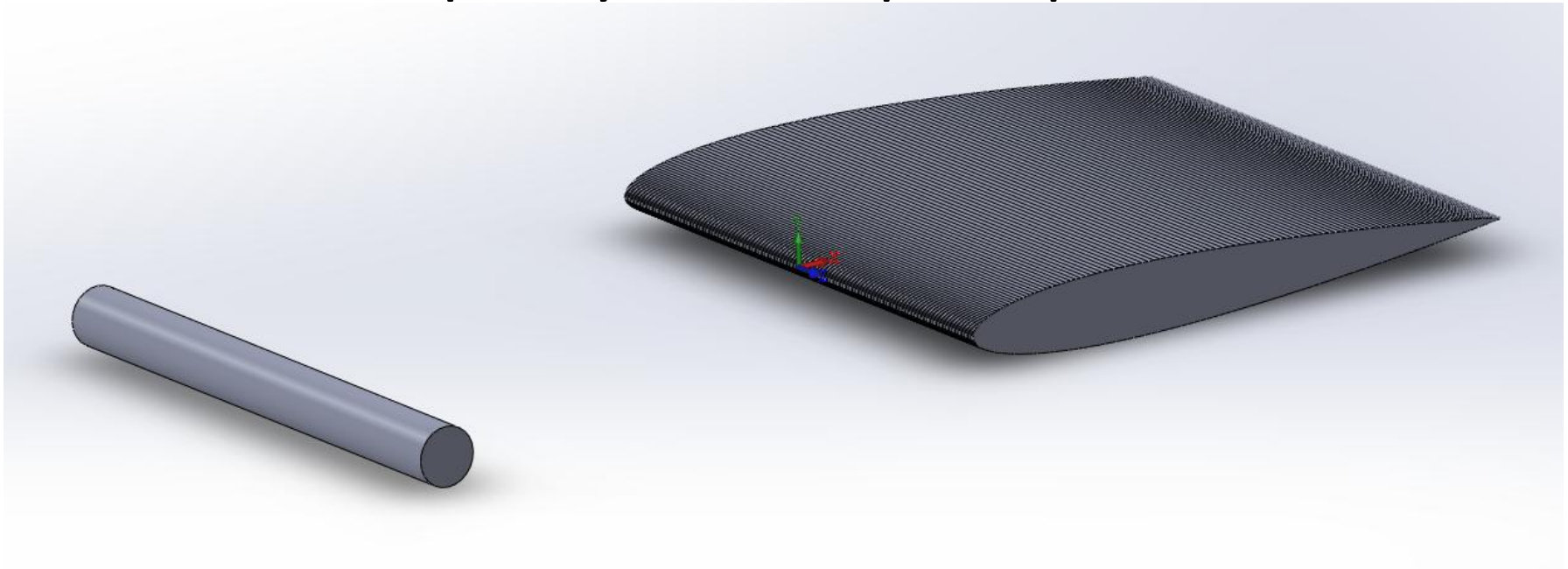
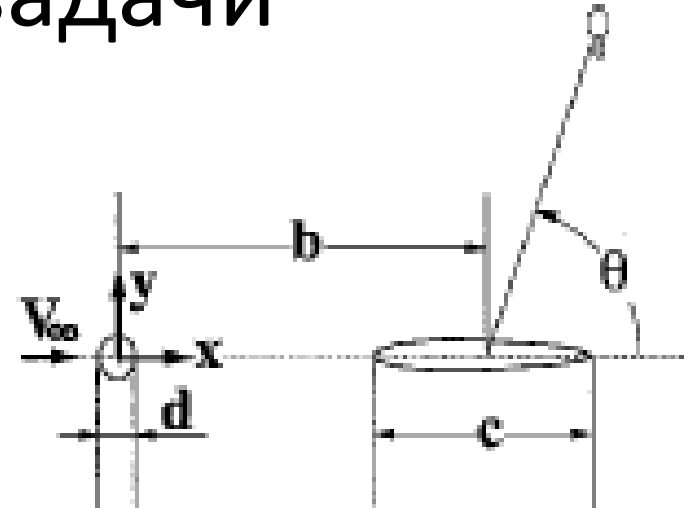
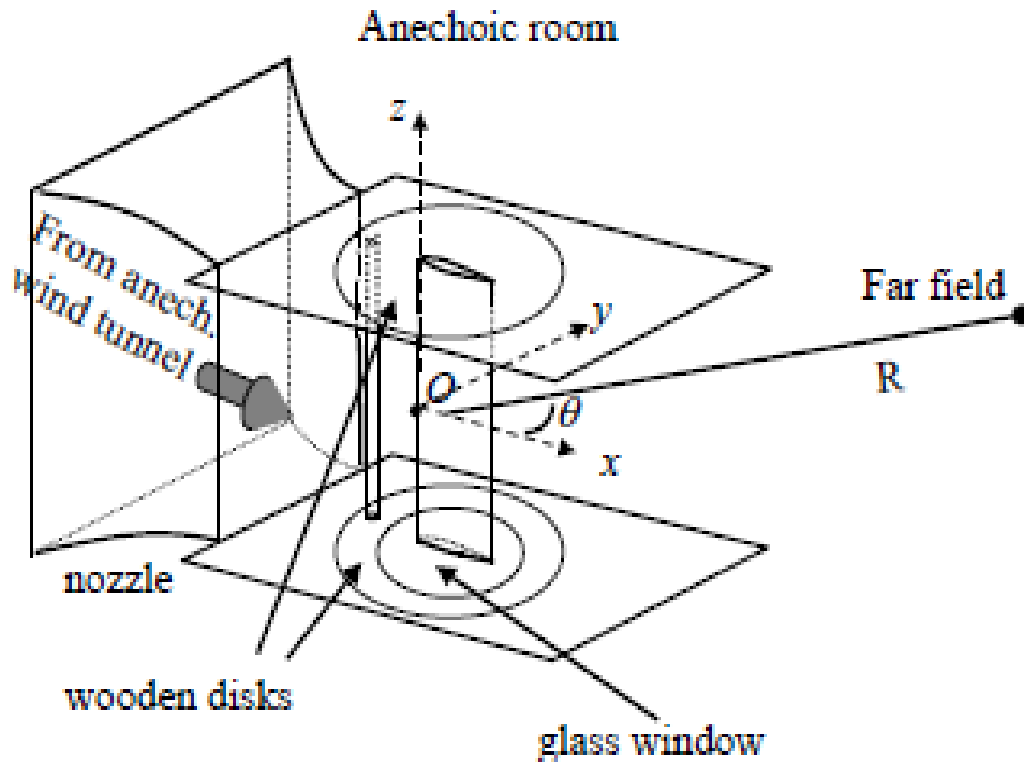


Рис. Цилиндр и профиль крыла NACA0012

- 1) Построение STL поверхности
- 2) Построение сетки (Salome + snappyHexMesh)
- 3) Расчет (OpenFOAM, pisoFoam)
- 4) Визуализация - Paraview

Постановка задачи



$$c=0.1 \text{ м}$$

$$d=0.1c=0.01 \text{ м}$$

$$V=72 \text{ м/с}$$

$$\rho=1.2 \text{ кг/м}^3$$

$$\mu=1.5 \cdot 10^{-5}$$

$$Re_d=4.8 \cdot 10^4$$

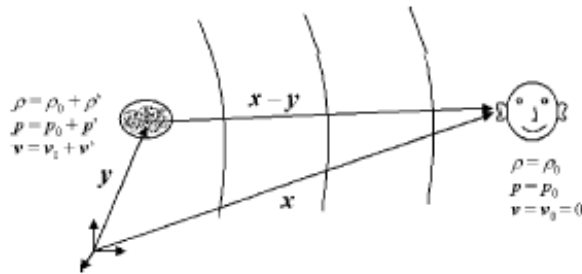
$$Re_c=4.8 \cdot 10^5$$

Рис. Схема эксперимента с PIV

$$z=3d=0.03 \text{ м}$$

Jacob M. et al. A rod-airfoil experiment as benchmark for broadband noise modeling. J. Theoret. Comput. Fluid Dynamics. 2005. Vol 19. pp.171–196.

Акустические аналогии. Аналогия Лайтхилла (1952)



$$\frac{\partial^2 \rho'}{\partial t^2} - c_0^2 \frac{\partial \rho'}{\partial x_i} = \frac{\partial^2 T_{ij}}{\partial x_i \partial x_j} + \frac{\partial^2 \beta \rho f}{\partial t^2} - \frac{\partial f_i}{\partial x_i}$$

$$T_{ij} = \rho v_i v_j - \tau_{ij} + (p' - c_0^2 \rho') \delta_{ij}$$

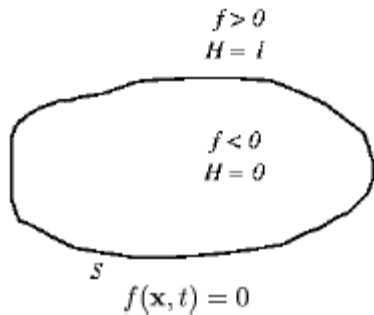
Аналогия Керла (1955)

$$\begin{aligned}
 (\rho - \rho_0)H(f) = & \frac{\partial^2}{\partial x_i \partial x_j} \int_V [T_{ij}] \frac{d^3 \mathbf{y}}{4\pi c_0^2 |\mathbf{x} - \mathbf{y}|} \\
 & - \frac{\partial}{\partial x_i} \int_S [\rho v_i v_j + (p - p_0) \delta_{ij} - \sigma_{ij}] \frac{dS_j(\mathbf{y})}{4\pi c_0^2 |\mathbf{x} - \mathbf{y}|} \\
 & + \frac{\partial}{\partial t} \int_S [\rho v_j] \frac{dS_j(\mathbf{y})}{4\pi c_0^2 |\mathbf{x} - \mathbf{y}|}
 \end{aligned}$$

$$4\pi c_0^2 (\rho(\tilde{\mathbf{x}}, t) - \rho_0) = -\frac{\partial}{\partial x_i} \left[\frac{F_i}{r} \right] = \frac{1}{c_0} \frac{x_i}{r^2} \left[\frac{\partial F_i}{\partial t} \right]$$

$$p'(\tilde{\mathbf{x}}, t) = \frac{1}{4\pi c_0} \frac{x_i}{r^2} \left[\frac{\partial F_i}{\partial t} + \frac{c_0 F_i}{r} \right]$$

Ffowks-Williams – Hawkins аналогия (1969)



$$\begin{aligned}
 & \left(\frac{1}{c_0^2} \frac{\partial^2}{\partial t^2} - \nabla^2 \right) (c_0^2 (\rho - \rho_0) H(f)) \quad \text{Dipole type term} \\
 & = \frac{\partial^2}{\partial x_i \partial x_j} (T_{ij} H(f)) - \frac{\partial}{\partial x_i} \left((\rho v_i (v_j - V_j) + (p - p_0) \delta_{ij} - \sigma_{ij}) \frac{\partial H}{\partial x_j} \right) \\
 & \quad \text{Quadrupole} \quad + \frac{\partial}{\partial t} \left((\rho (v_j - V_j) + \rho_0 V_j) \frac{\partial H}{\partial x_j} \right) \\
 & \quad \text{Monopole type term}
 \end{aligned}$$

Разработка библиотеки libAcoustics

Необходимо знать:

- 1) Какие данные необходимо собрать из CFD расчета
- 2) Как рассчитать поверхностный интеграл для определения обобщенной аэродинамической силы
- 3) Как рассчитать производную по времени для силы
- 4) Как определить акустического давление, значение SPL
- 5) Как сохранять данные на диске
- 6) Как организовать расчет в параллельном режиме

Библиотека **libAcoustics.so** в OpenFoam

содержит следующие файлы: Curle.C CurleFunctionObject.C
CurleFunctionObject.H Curle.H SoundObserver.C
SoundObserver.H FoamFftwDriver.C FoamFftwDriver.H

Результаты расчета. Сравнение с экспериментом

Rod results

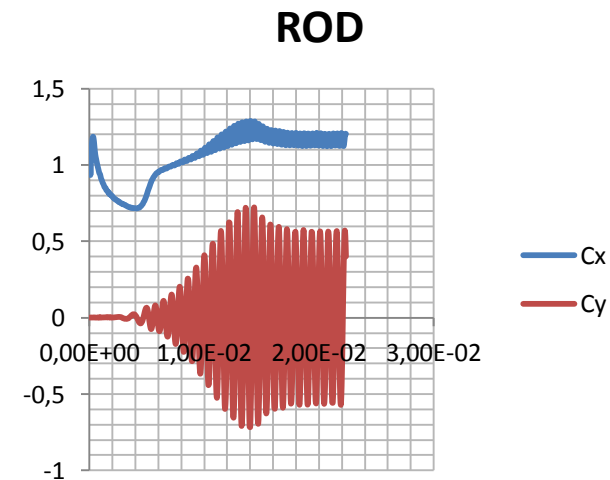
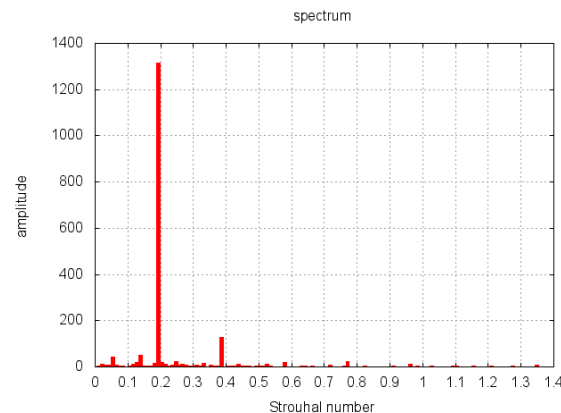
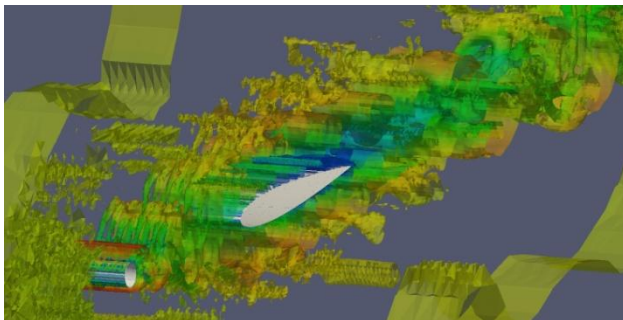
	U-RANS (rod-airfoil)	LES (rod-airfoil)	Measurements
St	0.24	0.19	0.19
$\langle C_D \rangle$	0.79 – 1.03	1.17	[1.0 ; 1.35]
C_D'	0.012 – 0.079	0.12	[0.08 ; 0.16]
C_L'	0.34 – 0.76	0.57	[0.4 ; 0.8]

$$St = 0.194$$

$$\langle C_D \rangle = 1.17$$

$$\Delta C_D = 0.08$$

$$\Delta C_L = 0.57$$



Расчет SPL с использованием библиотеки “libAcoustics”

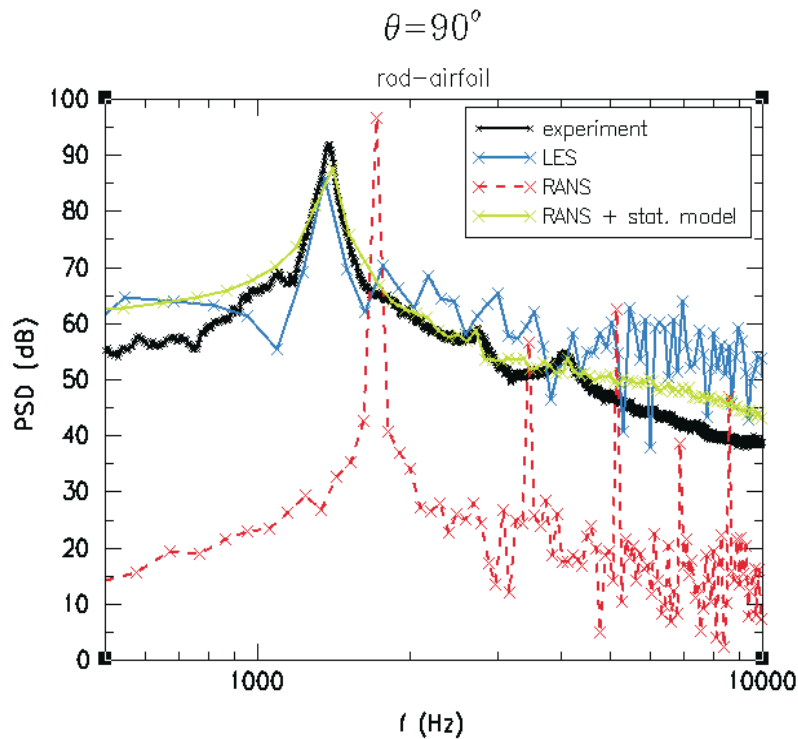


Рис. Результаты эксперимента

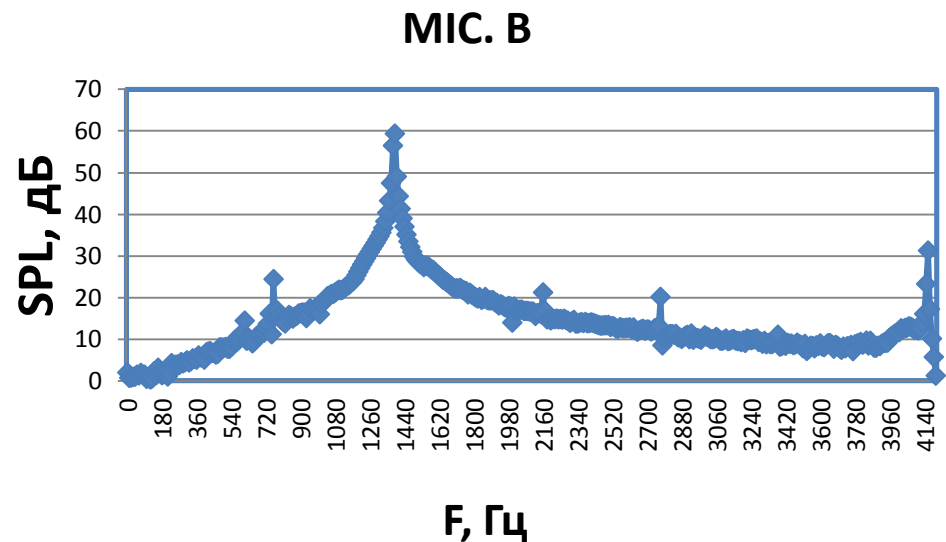
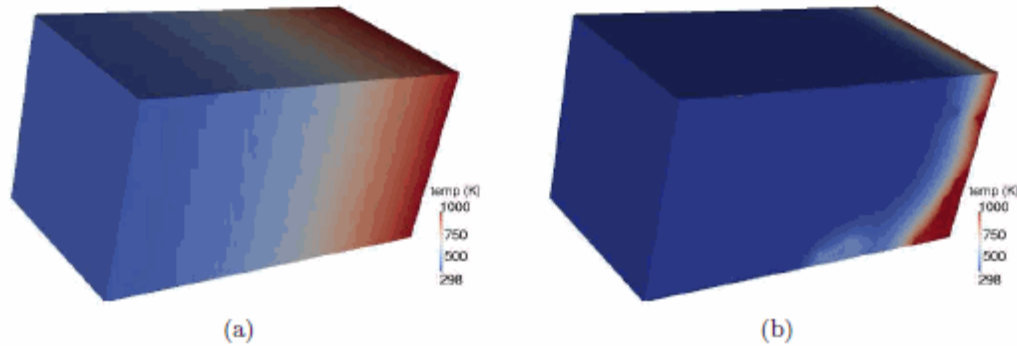


Рис. Результаты расчета SPL для микрофона В

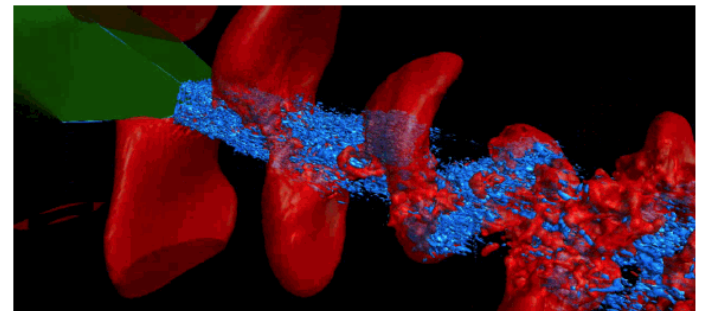
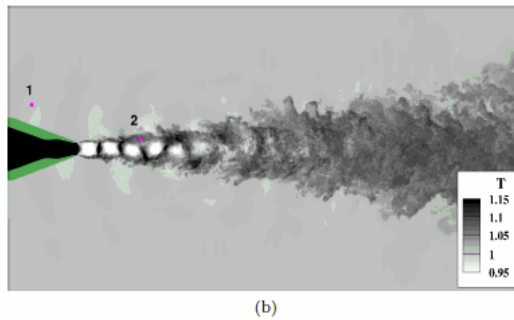
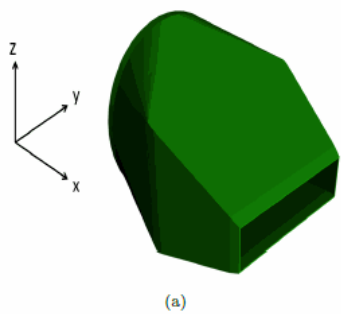
Задачи теплопроводности и аэроакустики



Матрица данных: 16 столбцов (радиус пузырька) и 4926801024 строк (257 x узлов x 129 y узлов x 129 z узлов x 9 times x 128 положений пузырька) > **600 GB**

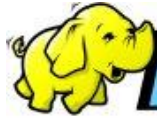
Матрица данных: 64 столбцов (радиус пузырька) и 4926801024 строк (257 x узлов x 129 y узлов x 129 z узлов x 9 times x 128 (положение пузырька)) > **2.2 TB 8192 расчетов**

Задача определения профиля температуры на выходе расчетной области в зависимости от радиуса пузырька



Когерентные структуры. Joseph W. Nichols et al. Stanford University. 2011-2013.

LES расчет на сетке 45 млн. ячеек с использованием более 1.5 млн. ядер 256 временных срезов. Каждый имеет 256 МВ. **Задача нахождения частот и амплитуд.**



hadoop Что такое Hadoop?

- Инфраструктура для параллельной обработки больших объемов данных (терабайты и петабайты)
- Особенности:
 - Функциональное программирование
 - Автоматическое распараллеливание
 - Перемещение вычислений к данным
- Свободно распространяемый набор утилит, библиотек и программный каркас для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов
- Open Source, <http://hadoop.apache.org>
- Основатели проекта компании Google (2004) и Yahoo! (2006)
- Учебные курсы: ИСП РАН и ИММ УРО РАН

Состав Hadoop

- Hadoop Common – общие компоненты Hadoop
- Hadoop HDFS – распределенная файловая система
- Hadoop MapReduce – модель программирования для обработки данных (Java, Ruby, Python, C++)
- Pig – язык для описания потока данных
- Hive – система для организации хранилищ данных
- HBase – распределенная столбцово-ориентированная база данных



HP AppSystem- Hadoop Appliance Cluster Architecture

Полная стойка (18 узлов) или Половина стойки (9 узлов)



2 Top of Rack Switches
5830AF-48G

2 Hadoop Masters
DL360p Gen8
RHEL 6.2, JobTracker, NameNode, Secondary NameNode (CDH4), BIOS Setup for CMU, NFS Server/Client, Sun Java 1.6.0_31, CM Agent

E5-2667 (2.90GHz/6-core/15MB/130W)
64GB (8x 8GB 1600MHz DDR3)
4x 900GB SAS 10K SFF Ent
4x 1GbE FlexLOM
P420i Array Controller (R1+0)

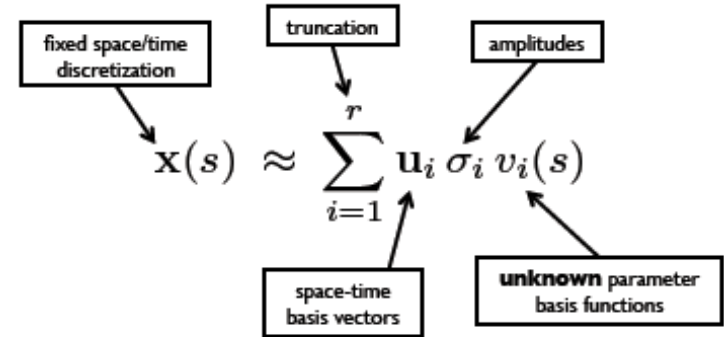
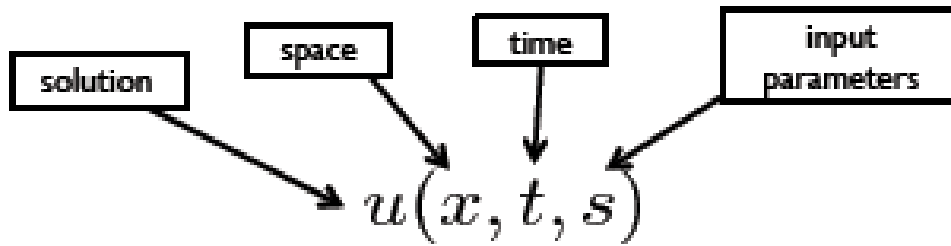
1 Management/Client/Edge Node
DL360pGen8
RHEL 6.2, HP Insight CMU 7.0, Cloudera Hue Server, Cloudera Manager4 (PostgreSQL), CDH4, NFS Server, Client Gateway for Shell Access (Pig and Hive installed), Sun Java 1.6.0_31

E5-2667 (2.90GHz/6-core/15MB/130W)
64GB (8x 8GB 1600MHz DDR3)
8x 900GB SAS 10K SFF Ent
4x 1GbE FlexLOM
P420i Array Controller (R1+0 boot;R5 rest)

9 or 18 Hadoop Data Nodes
DL380e Gen8
RHEL 6.2, TaskTracker, DataNode (CDH4), BIOS Setup for CMU, Sun Java 1.6.0_31, CM Agent, mirrored boot volumes, R0 on the rest

E5-2440 (2.40GHz/6-core/15MB/95W)
48GB (6x 8GB 1333MHz DDR)
2x 500GB SATA 7.2K LFF MDL
12x 2TB SATA 7.2K LFF MDL
4x 1GbE FlexLOM
P420 Array Controller (R1+0 boot ONLY – not data; JBOD/R0 rest)

Математические алгоритмы для обработки данных (SVD, POD, DMD)

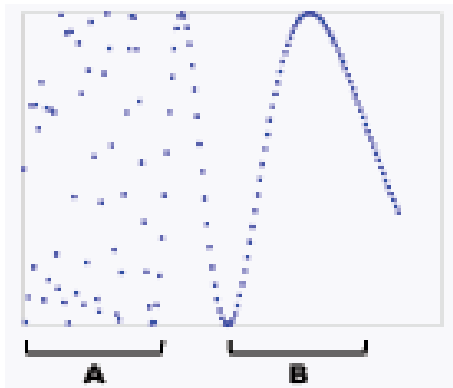


Proper orthogonal decomposition (POD) modes

experimental design

$$\begin{bmatrix} x(s_1) & \dots & x(s_n) \end{bmatrix} = X \xrightarrow{\text{SVD}} U \Sigma V^T$$

$$V_{ij} = v_i(s_j) \quad \{v_i(s_j)\} \longrightarrow v_i(s)$$



CAQR (Demmel et al, 2008)

$$\begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \end{bmatrix} = \underbrace{\begin{bmatrix} Q_1 & & & \\ & Q_2 & & \\ & & Q_3 & \\ & & & Q_4 \end{bmatrix}}_{8n \times 4n} \underbrace{\begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \end{bmatrix}}_{4n \times n}$$

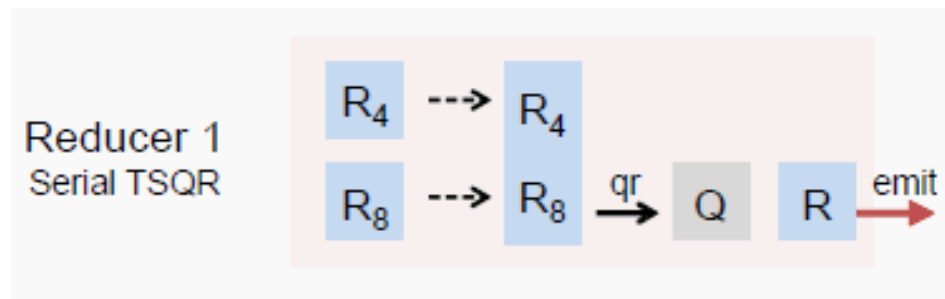
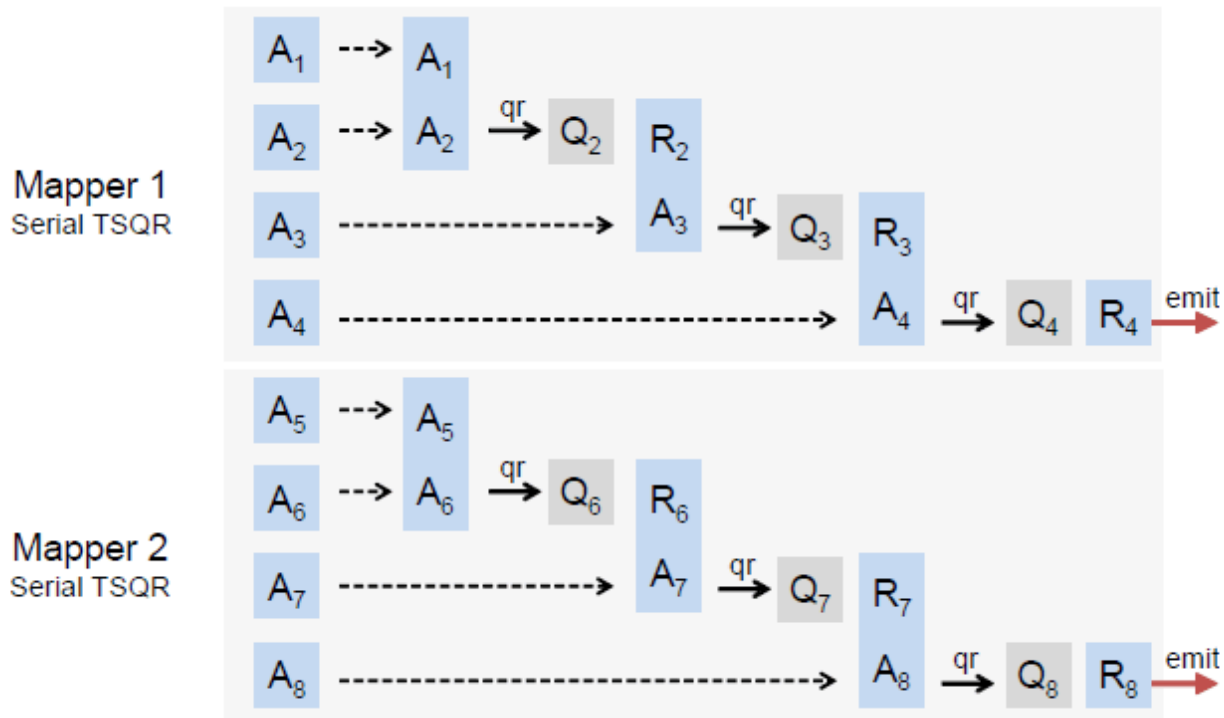
$$= \underbrace{\begin{bmatrix} Q_1 & & & \\ & Q_2 & & \\ & & Q_3 & \\ & & & Q_4 \end{bmatrix}}_{8n \times 4n} \underbrace{\begin{bmatrix} \tilde{Q} \\ \tilde{R} \end{bmatrix}}_{4n \times n \quad n \times n}$$



<http://github.com/dgleioh/nrtsqr>

Базовые функции и интерполяционные методы

Алгоритм для матрицы A (8 столбцов)



Dynamic Mode Decomposition

$$\{\psi_0, \psi_1, \dots, \psi_N\},$$

$$\Psi_0^{N-1} = [\psi_0 \ \psi_1 \ \dots \ \psi_{N-1}] \in \mathbb{C}^{M \times N},$$

$$\Psi_1^N = [\psi_1 \ \psi_2 \ \dots \ \psi_N] \in \mathbb{C}^{M \times N}$$

$$\begin{aligned} \Psi_1^N &= [\psi_1 \ \psi_2 \ \dots \ \psi_N] \\ &= [A\psi_0 \ A\psi_1 \ \dots \ A\psi_{N-1}] \\ &= A\Psi_0^{N-1}. \end{aligned}$$

$$F = \begin{bmatrix} f(x_1, s_1) & \dots & f(x_1, s_N) \\ \vdots & \ddots & \vdots \\ f(x_M, s_1) & \dots & f(x_M, s_N) \end{bmatrix}.$$

$$\psi_t = \sum_{i=1}^N \beta_i \lambda_i^t \theta_i, \quad t \in \{0, \dots, N-1\}.$$

$$\text{minimize}_{\alpha} J(\alpha) = \|\Sigma V^* - Y D_{\alpha} V_{\text{and}}\|_F^2,$$

Модель редукции в MapReduce

- Создать Reduced order model, результаты вычислений пересылаются и хранятся на Hadoop кластере.

Основные этапы:

- 1. Сформировать плотную матрицу F исходя из данных вычислений
- 2. Рассчитать единственное значение разложения матрицы F
- 3. Рассчитать значения искомой функции, определить характерные частоты и моды

TSQR и SVD in Hadoop

- Tall-and-skinny QR (TSQR) разложение матрицы
- Разбить Tall-and-skinny матрицу F на множество маленьких матриц которые должны быть декомпозированы независимо с помощью QR разложения
- Hadoop Distributed File System хранит матрицу в форме небольших “chunks” в соответствии с внутренней процедурой расщепления
- Каждая map function читает маленькую локальную матрицу и рассчитывает QR разложение. Далее записывает Q factor на диск с малом tag ϕ .

Заключение

- В вычислительной аэрогидродинамике необходимо проводить большое количество расчетов
- Пример: “rod-airfoil”, расчет значения SPL
- Влияние изменения параметров на определяющие значения функций (SPL, температура, давление)
- Необходим эффективный алгоритм для анализа больших данных
- Hadoop/Mapreduce – эффективный инструмент
- POD и Dynamic Mode Decomposition алгоритмы
- QR алгоритм для решения задачи о собственных значениях (частот и амплитуд) в реализации Mapreduce (скрипты)
- Проект на 2014 в рамках группы UniCFD и web-лаборатории UniHUB