SAND2015-9290C

Illustration by Chris Brigman

Parallel Tucker Compression for Large-scale Scientific Data

Woody Austin
Univ. Texas, Austin, TX

Grey Ballard and Tamara G. Kolda*
Sandia National Laboratories, Livermore, CA

SIAM Conference on Applied Linear Algebra (AL15)
Atlanta, Georgia, USA
October 29, 2015

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

A Tensor is an N-Way Array

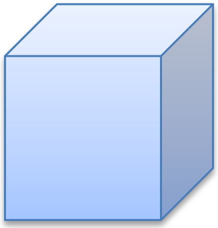
Vector
N = 1

$\mathbf{x}$

Matrix
N = 2

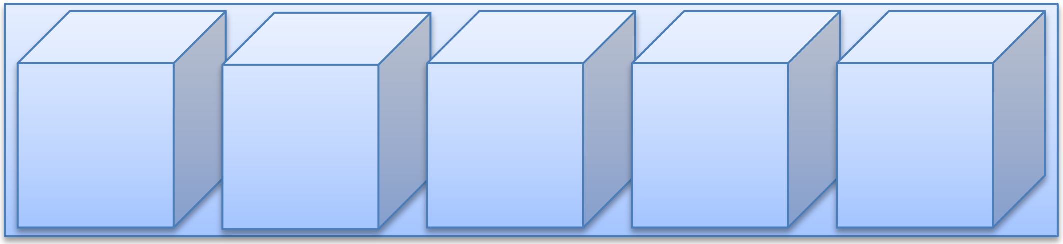
$\mathbf{X}$

3rd-Order Tensor
N = 3



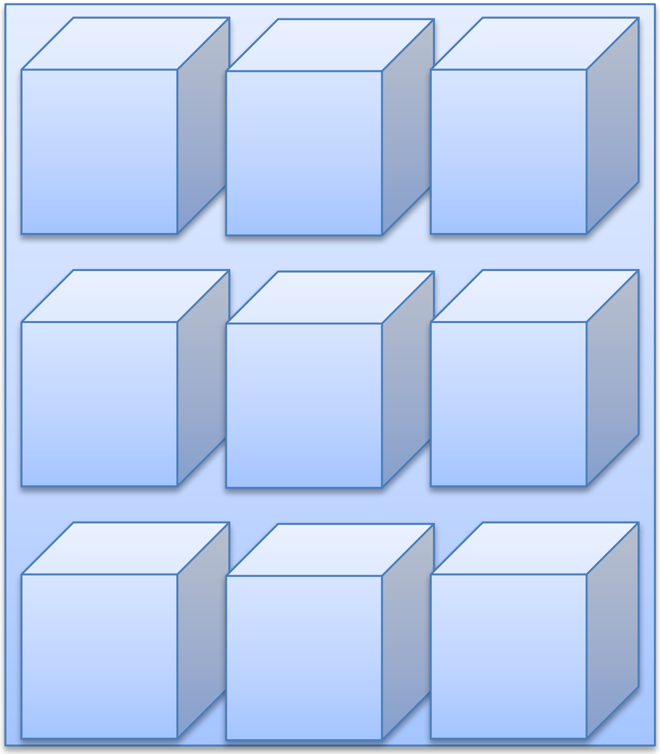
$\mathcal{X}$

4th-Order Tensor
N = 4



$\mathcal{X}$

5th-Order Tensor
N = 5

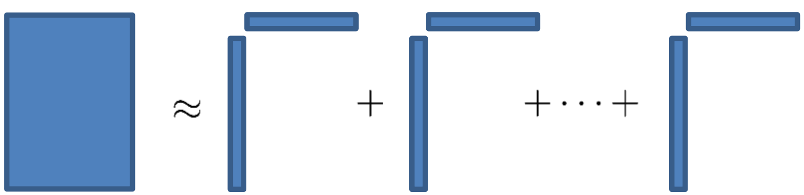


$\mathcal{X}$

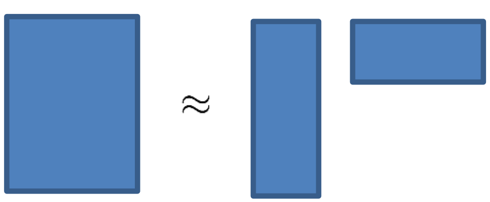
Tensor Decompositions are the New Matrix Decompositions

Singular value decomposition (SVD), eigendecomposition (EVD), nonnegative matrix factorization (NMF), sparse SVD, etc.

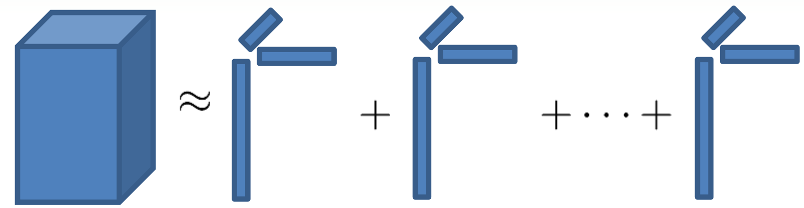
Viewpoint 1: Sum of outer products, useful for interpretation



Viewpoint 2: High-variance subspaces, useful for compression

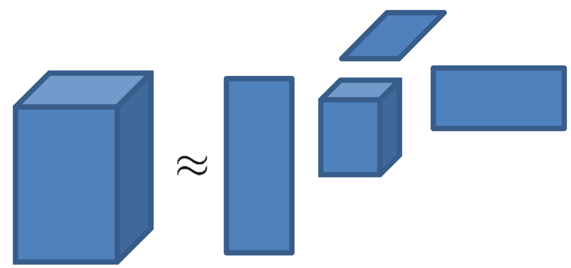


CP Model: Sum of d-way outer products, useful for interpretation



CANDECOMP, PARAFAC, Canonical Polyadic, CP

Tucker Model: Project onto high-variance subspaces to reduce dimensionality



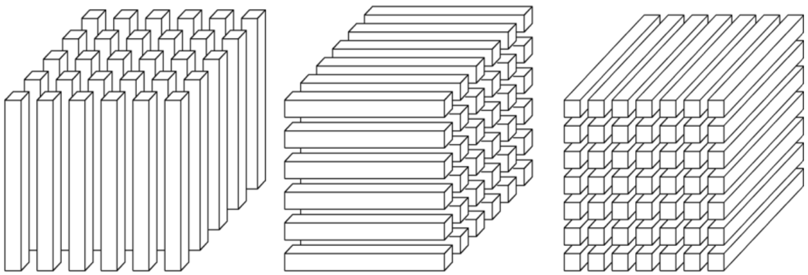
HOSVD, Best Rank-(R1,R2,…,RN) decomposition

Other models for compression include hierarchical Tucker and tensor train.

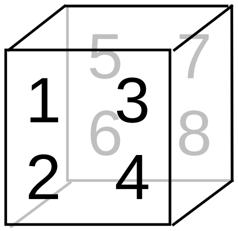
Tensor Fibers, Mode-n Unfolding, and Mode-n Multiplication

Sandia National Laboratories

Tensor "mode-n fibers" analogous to matrix rows and columns



$\mathbf{X}_{(n)}$ denotes mode-n unfolding, arranges mode-n fibers as matrix columns

$\mathcal{X} =$ 

$$\mathbf{X}_{(1)} = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 2 & 4 & 6 & 8 \end{bmatrix}$$

$$\mathbf{X}_{(2)} = \begin{bmatrix} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \end{bmatrix}$$

$$\mathbf{X}_{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}$$

$$n\text{-rank}(\mathcal{X}) = \text{col-rank}(\mathbf{X}_{(n)})$$

Tensor-times-matrix (TTM) in mode-n multiplies mode-n fibers times matrix

$$I_1 \times \cdots \times I_n \times \cdots \times I_N \qquad K \times I_n$$

$$\boxed{\mathcal{Y} = \mathcal{X} \times_n \mathbf{U}}$$

$$I_1 \times \cdots \times K \times \cdots \times I_N$$

Equivalent to matrix operation:

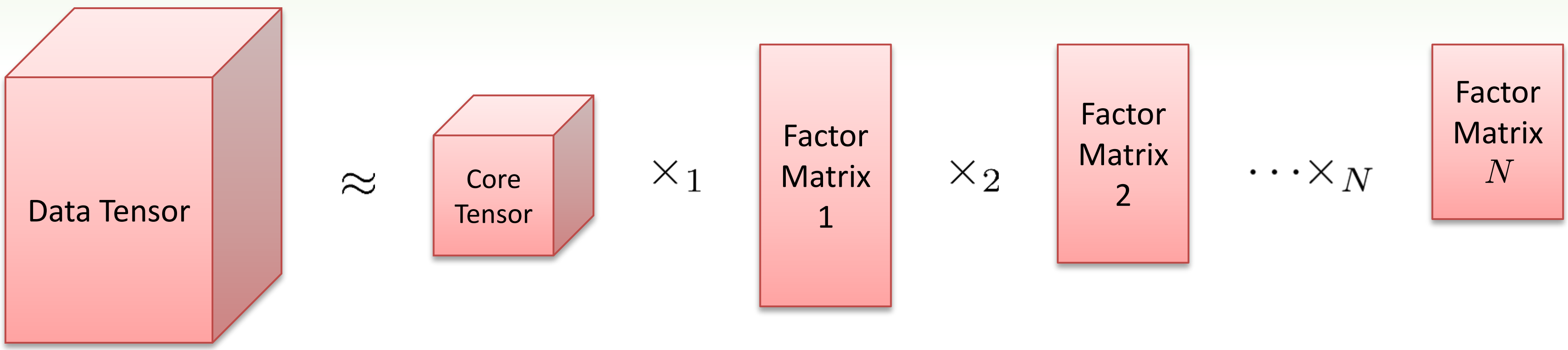
$$K \times I_n$$

$$K \times \hat{I}_n \rightarrow \boxed{\mathbf{Y}_{(n)} = \mathbf{U}\mathbf{X}_{(n)}} \qquad I_n \times \hat{I}_n$$

$$I = \prod I_n, \quad \hat{I}_n = I / I_n$$

Tucker Decomposition

$$\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)}$$



$$I_1 \times I_2 \times \cdots \times I_N \qquad R_1 \times R_2 \times \cdots \times R_N \qquad I_1 \times R_1 \qquad I_2 \times R_2 \qquad I_N \times R_N$$

$$\min_{\hat{\mathcal{X}}} \sum_{i_1 \dots i_N} (x_{i_1 \dots i_N} - \hat{x}_{i_1 \dots i_N})^2 \text{ subject to } \hat{\mathcal{X}} = \mathcal{G} \times \{ \mathbf{U}^{(n)} \}$$

WLOG, assume $\mathbf{U}^{(n)}$ has orthogonal columns for all n.

If $R_n \geq$ rank($\mathbf{X}_{(n)}$) for all n, then decomposition is exact. Else, it's lossy.

Tucker (1966); Kapteyn, Neudecker, Wansbeek (1986)

Optimization Problem

$$\min_{\hat{\mathcal{X}}} \sum_{i_1 \ldots i_N} (x_{i_1 \ldots i_N} - \hat{x}_{i_1 \ldots i_N})^2 \text{ subject to } \hat{\mathcal{X}} = \mathcal{G} \times \{ \mathbf{U}^{(n)} \}$$

<u>Homework</u>: (1) At an optimum, it must be the case that

$$\mathcal{G} = \mathcal{X} \times \{ \mathbf{U}^{(n)\mathsf{T}} \}$$

(2) The minimization problem above can be written as

$$\max \sum_{i_1 \ldots i_N} z_{i_1 \ldots i_N}^2 \text{ subject to } \mathcal{Z} = \mathcal{X} \times \{ \mathbf{U}^{(n)\mathsf{T}} \}$$

$$\max \| \mathbf{U}^{(n)\mathsf{T}} \mathbf{W}_{(n)} \|_F^2 \text{ subject to } \mathcal{W} = \mathcal{X} \times \{ \mathbf{U}^{(m)\mathsf{T}} \}_{m \neq n}$$

Solution to (*) is to choose $\mathbf{U}^{(n)}$ has to be the R_n leading left singular vectors of $\mathbf{W}_{(n)}$.

Truncated HOSVD Chooses Ranks for Requested Error

procedure T-HOSVD($\mathcal{X}$, ϵ)
  for $n = 1, \ldots, N$ **do**
    $\mathbf{S}^{(n)} \leftarrow \mathbf{X}_{(n)} \mathbf{X}_{(n)}^{\mathsf{T}}$
    $R_n \leftarrow \min R$ such that $\sum_{i>R} \lambda_i(\mathbf{S}^{(n)}) \leq \epsilon^2 \|\mathcal{X}\|^2 / N$
    $\mathbf{U}^{(n)} \leftarrow$ leading R_n eigenvectors of $\mathbf{S}^{(n)}$
  end for
  $\mathcal{G} \leftarrow \mathcal{X} \times \{ \mathbf{U}^{(n)\mathsf{T}} \}$
  return $(\mathcal{G}, \{ \mathbf{U}^{(n)} \})$
end procedure

Also known as “Tucker1” method.

$$\|\mathcal{X} - \hat{\mathcal{X}}\|^2 \leq \sum_{n=1}^{N} \left(\sum_{i=R_n+1}^{I_n} \lambda_i(\mathbf{S}^{(n)}) \right) \leq \epsilon \|\mathcal{X}\|$$

Tucker (1966); De Lathauwer, De Moor, Vandewalle (2000); Vannieuwenhoven, Vandebril, and Meerbergen (2012)

Sequentially Truncated HOSVD improves further

procedure ST-HOSVD($\mathcal{X}$, ϵ)
  $\mathcal{Y} \leftarrow \mathcal{X}$
  for $n = 1, \ldots, N$ **do**
    $\mathbf{S}^{(n)} \leftarrow \mathbf{Y}_{(n)} \mathbf{Y}_{(n)}^{\mathsf{T}}$
    $R_n \leftarrow \min R$ such that $\sum_{r>R} \lambda_r(\mathbf{S}^{(n)}) \leq \epsilon^2 \|\mathcal{X}\|^2 / N$
    $\mathbf{U}^{(n)} \leftarrow$ leading R_n eigenvectors of $\mathbf{S}^{(n)}$
    $\mathcal{Y} \leftarrow \mathcal{Y} \times_n \mathbf{U}^{(n)\mathsf{T}}$
  end for
  $\mathcal{G} \leftarrow \mathcal{Y}$
  return $(\mathcal{G}, \{\, \mathbf{U}^{(n)} \,\})$
end procedure

Smaller at each step.

$$\|\mathcal{X} - \hat{\mathcal{X}}\|^2 = \sum_{n=1}^{N} \left(\sum_{i=R_n+1}^{I_n} \lambda_i(\mathbf{S}^{(n)}) \right) \leq \epsilon \|\mathcal{X}\|$$

Vannieuwenhoven, Vandebril, and Meerbergen (2012)

Higher-Order Orthogonal Iteration (HOOI) improves again

$$
\begin{array}{l}
\textbf{procedure } \mathrm{HOOI}(\mathcal{X}, \epsilon) \\
\quad (\mathcal{G}, \{\, \mathbf{U}^{(n)} \,\}) = \text{ST-HOSVD}(\mathcal{X}, \epsilon) \\
\quad \textbf{repeat} \\
\qquad \textbf{for } n = 1, \ldots, N \textbf{ do} \\
\qquad\quad \mathcal{Y} \leftarrow \mathcal{X} \times \{\, \mathbf{U}^{(m)\mathsf{T}} \,\}_{m \neq n} \\
\qquad\quad \mathbf{S}^{(n)} \leftarrow \mathbf{Y}_{(n)} \mathbf{Y}_{(n)}^{\mathsf{T}} \\
\qquad\quad \mathbf{U}^{(n)} \leftarrow \text{leading } R_n \text{ eigenvectors of } \mathbf{S}^{(n)} \\
\qquad \textbf{end for} \\
\qquad \mathcal{G} \leftarrow \mathcal{Y} \times_N \mathbf{U}^{(N)\mathsf{T}} \\
\quad \textbf{until} \text{ the quantity } (\|\mathcal{X}\|^2 - \|\mathcal{G}\|^2) \text{ ceases to decrease} \\
\quad \textbf{return } (\mathcal{G}, \{\, \mathbf{U}^{(n)} \,\}) \\
\textbf{end procedure}
\end{array}
$$

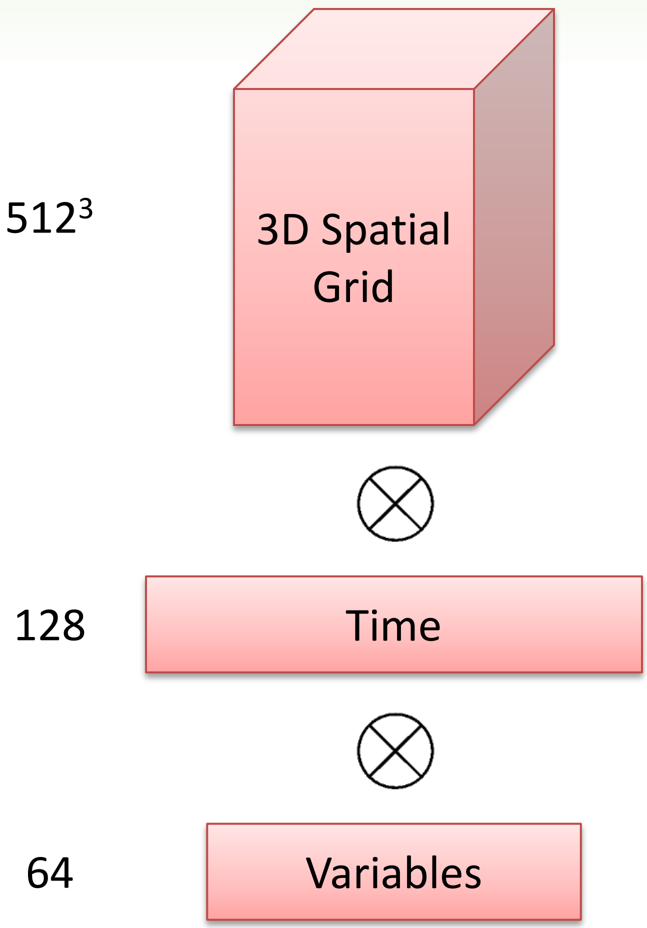
Kroonenberg and De Leeuw (1980); Kapteyn, Neudecker, Wansbeek (1986); De Lathauwer, De Moor, Vandewalle (2000)

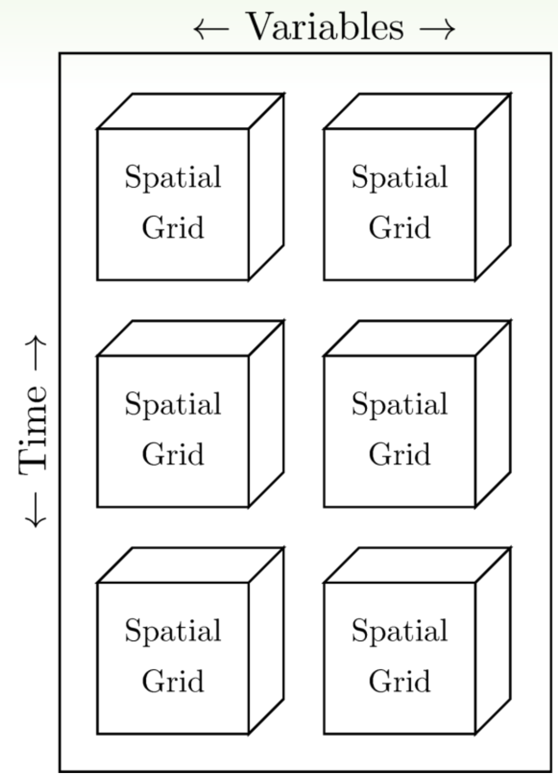
Key Kernels in ST-HOSVD are TTM and Gram

procedure ST-HOSVD($\mathcal{X}$, ϵ)
  $\mathcal{Y} \leftarrow \mathcal{X}$
  for $n = 1, \ldots, N$ **do**
Gram $\boxed{\mathbf{S}^{(n)} \leftarrow \mathbf{Y}_{(n)} \mathbf{Y}_{(n)}^{\mathsf{T}}}$
    $R_n \leftarrow \min R$ such that $\sum_{r>R} \lambda_r(\mathbf{S}^{(n)}) \leq \epsilon^2 \|\mathcal{X}\|^2 / N$
    $\mathbf{U}^{(n)} \leftarrow$ leading R_n eigenvectors of $\mathbf{S}^{(n)}$
TTM $\boxed{\mathcal{Y} \leftarrow \mathcal{Y} \times_n \mathbf{U}^{(n)\mathsf{T}}}$
  end for
  $\mathcal{G} \leftarrow \mathcal{Y}$
  return $(\mathcal{G}, \{ \mathbf{U}^{(n)} \})$
end procedure

Vannieuwenhoven, Vandebril, and Meerbergen (2012)

Tensors in Scientific Applications are Huge, Need Parallel Methods



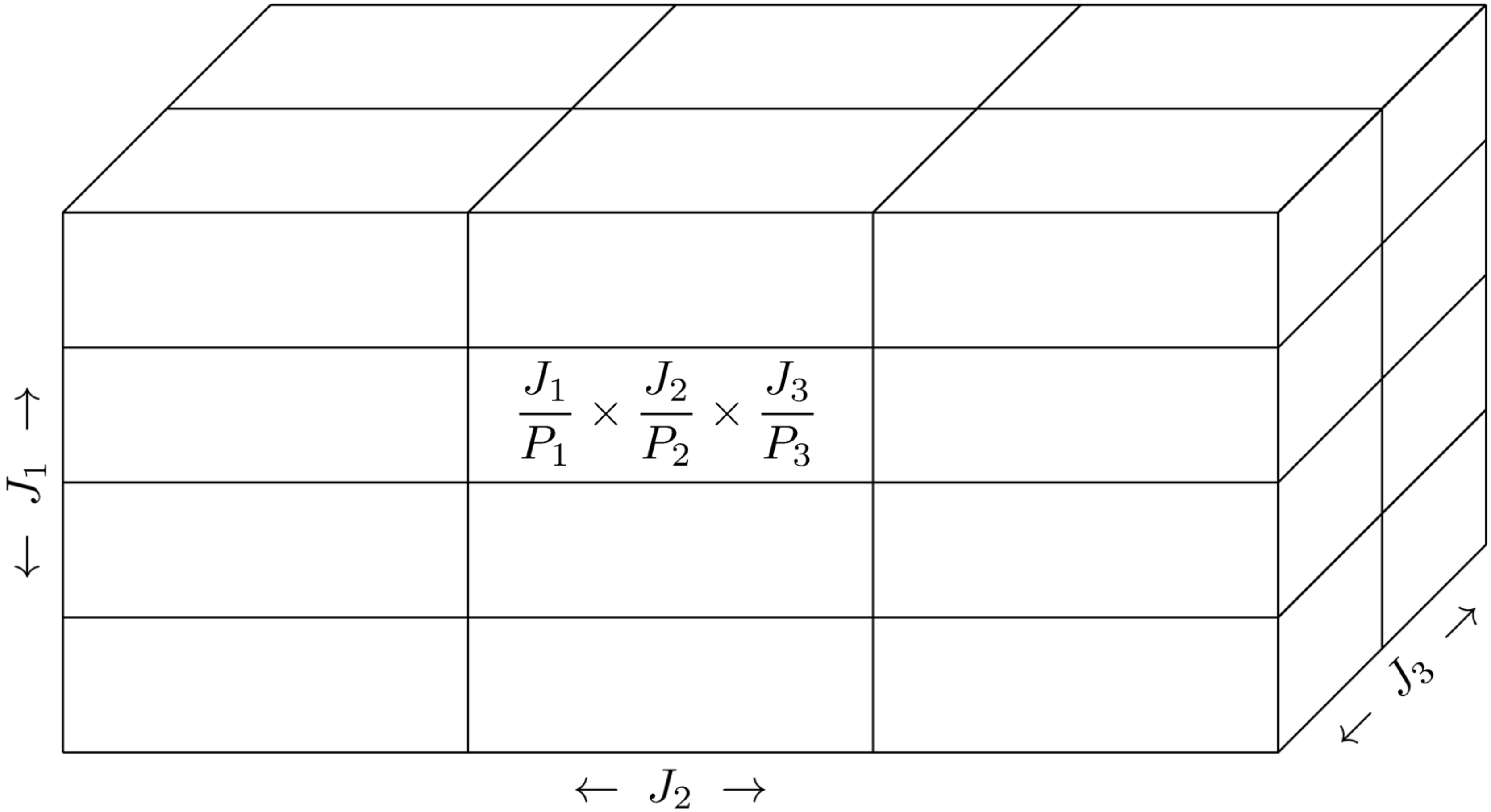


2^{40} elements

8TB (double precision)

Tensor Distribution: Cartesian

Processor Grid: $P_1 \times P_2 \times P_3$



Unfolded Tensor Distribution

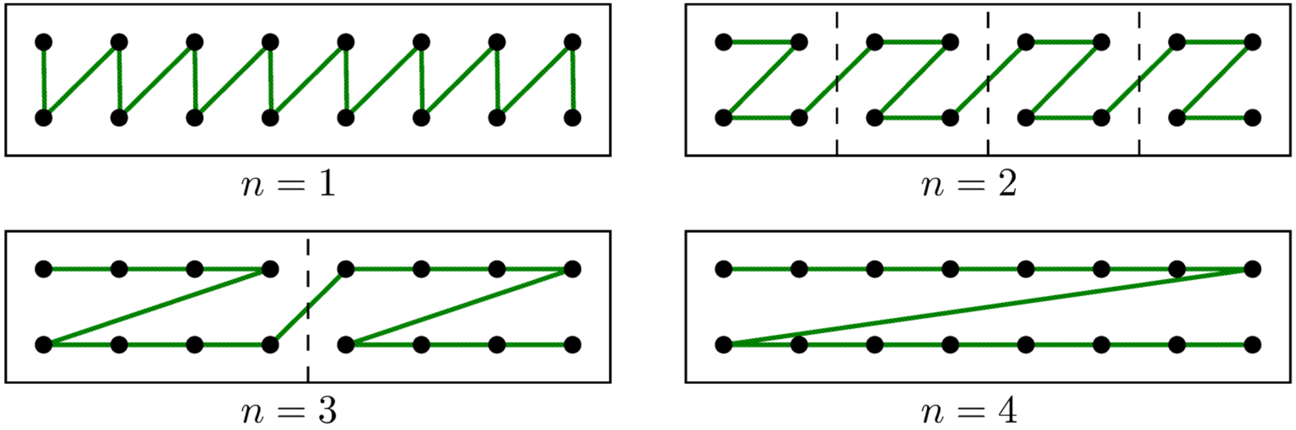
Global Tensor Size: $J_1 \times J_2 \times \cdots \times J_N, \quad J = \prod J_n, \quad \hat{J}_n = J/J_n$

Processor Grid Size: $P_1 \times P_2 \times \cdots \times P_N, \quad P = \prod P_n, \quad \hat{P}_n = P/P_n$

Global Unfolded Tensor: $J_n \times \hat{J}_n$

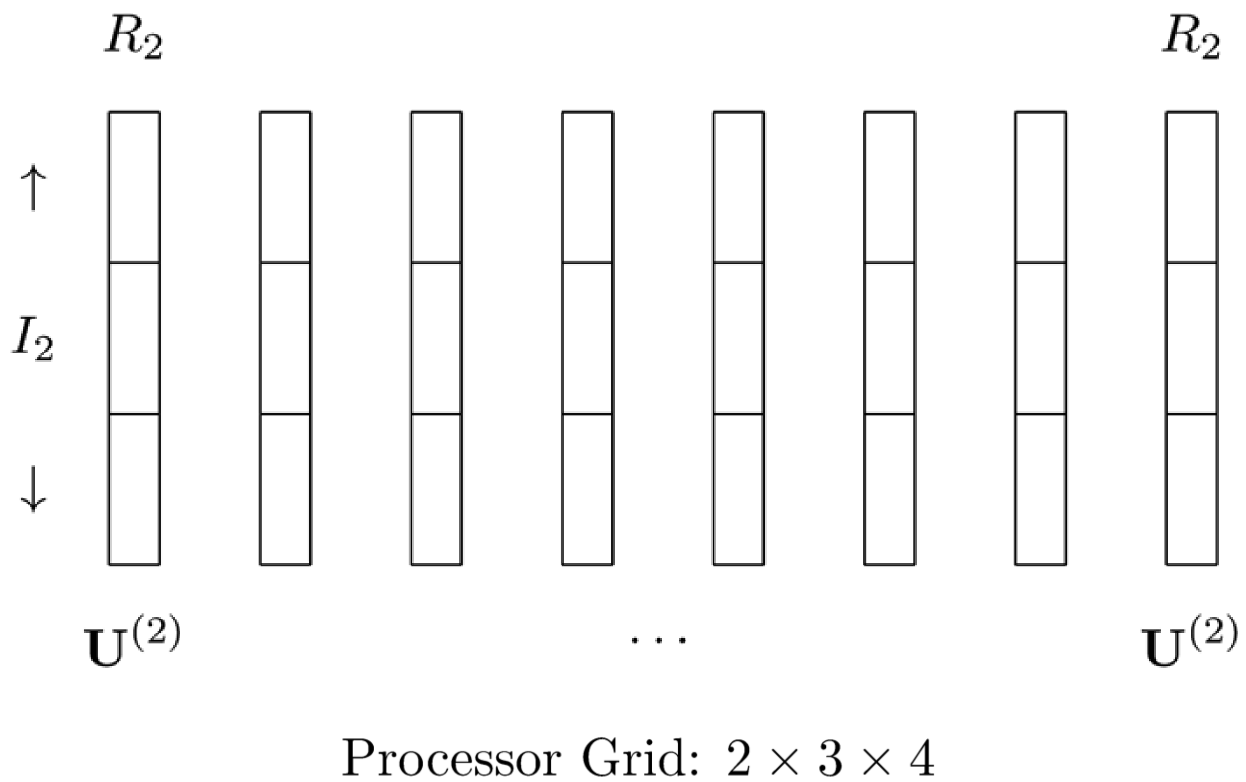
Processor Grid: $P_n \times \hat{P}_n$

Local Layout: 2 x 2 x 2 x 2



Redundant Factor Matrix Distribution

Factor matrices are replicated on each processor fiber and 1D row-distributed on each fiber

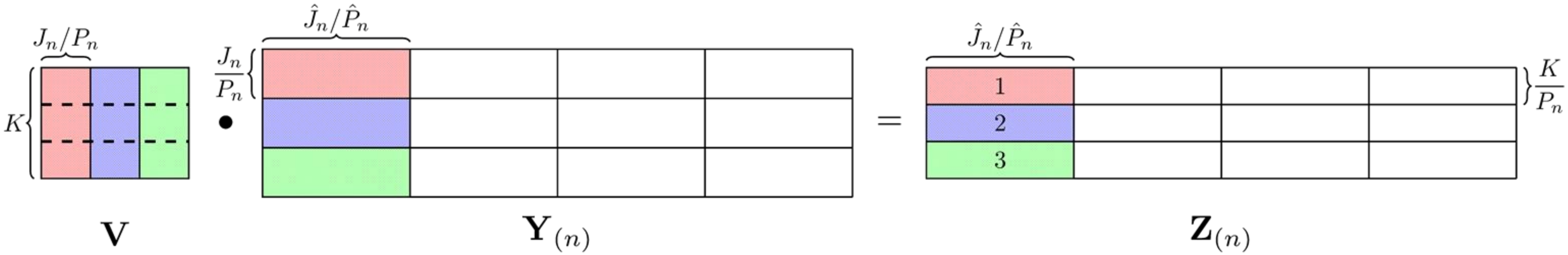


Parallel TTM "shrinks" the Tensor

$$\mathcal{Z} = \mathcal{Y} \times_n \mathbf{V} \quad \Leftrightarrow \quad \mathbf{Z}_{(n)} = \mathbf{V}\mathbf{Y}_{(n)}$$

Global Tensor Size: $J_1 \times J_2 \times \cdots \times J_N, \quad J = \prod J_n, \quad \hat{J}_n = J/J_n$

Processor Grid Size: $P_1 \times P_2 \times \cdots \times P_N, \quad P = \prod P_n, \quad \hat{P}_n = P/P_n$



procedure TTM($\mathcal{Y}$,$\mathbf{V}$,n)
  myProcID $\leftarrow (p_1, p_2, \ldots, p_N)$
  myProcCol $\leftarrow (p_1, \ldots, p_{n-1}, *, p_{n+1}, \ldots, p_N)$
  for $\ell = 1, \ldots, P_n$ **do**
    $\mathcal{W} \leftarrow \bar{\mathcal{Y}} \times_n \bar{\mathbf{V}}^{[\ell]}$
    $\bar{\mathcal{Z}} \leftarrow$ REDUCE($\mathcal{W}$, myProcCol, ℓ)
  end for
  return $\mathcal{Z}$
end procedure

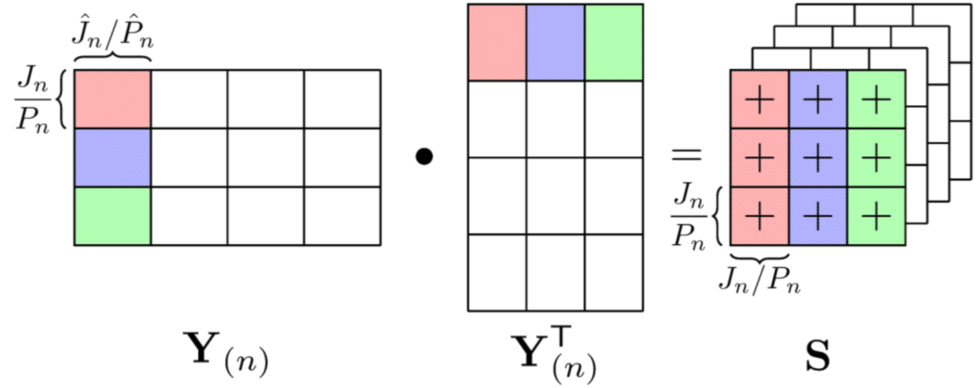
$$C_{\text{TTM}} = 2\gamma \frac{JK}{P} + \alpha P_n \log P_n + \beta (P_n - 1) \frac{\hat{J}_n K}{P}$$

$$M_{\text{TTM}} = \underbrace{J/P}_{\bar{\mathcal{Y}}} + \underbrace{J_n K/P_n}_{\bar{\mathbf{V}}} + \underbrace{\hat{J}_n K/P}_{\bar{\mathcal{Z}}} + \underbrace{\hat{J}_n K/P}_{\mathcal{W}}.$$

Parallel Gram

$$\mathbf{S} = \mathbf{Y}_{(n)} \mathbf{Y}_{(n)}^{\mathsf{T}}$$

procedure GRAM($\boldsymbol{\mathcal{Y}}$, n)
  myProcID $\leftarrow (p_1, p_2, \ldots, p_N)$
  myProcCol $\leftarrow (p_1, \ldots, p_{n-1}, *, p_{n+1}, \ldots, p_N)$
  myProcRow $\leftarrow (*, \ldots, *, p_k, *, \ldots, *)$
  $\mathbf{V}^{[p_n]} \leftarrow \bar{\mathbf{Y}}_{(n)} \bar{\mathbf{Y}}_{(n)}^{\mathsf{T}}$
  for $i = 1$ to $P_n - 1$ **do**
    $j \leftarrow (p_n - i) \mod P_n$
    $k \leftarrow (p_n + i) \mod P_n$
    Send $\bar{\boldsymbol{\mathcal{Y}}}$ to process $(p_1, \ldots, p_{n-1}, j, \ldots, p_N)$
    Receive $\boldsymbol{\mathcal{W}}$ from process $(p_1, \ldots, p_{n-1}, k, \ldots, p_N)$
    $\mathbf{V}^{[k]} \leftarrow \bar{\mathbf{Y}}_{(n)} \mathbf{W}_{(n)}^{\mathsf{T}}$
  end for
  $\bar{\mathbf{S}}$ = All-Reduce($\mathbf{V}$, myProcRow)
  return $\bar{\mathbf{S}}$
end procedure



$$\begin{aligned} C_{\text{GRAM}} =& \gamma 2 J_n J / P + 2(P_n - 1)\left(\alpha + \beta J / P\right) \\ & + 2\alpha \log \hat{P}_n + 2\beta(\hat{P}_n - 1) J_n^2 / P \\ M_{\text{GRAM}} =& J/P + J/P + J_n^2 / P_n + J_n^2 / P_n \end{aligned}$$

Application Results: Compression versus Accuracy

Ranks depend on error:

$$\|\mathcal{X} - \mathcal{M}\|^2 \leq \sum_{n=1}^{N} \left(\sum_{i=R_n+1}^{I_n} \lambda_i(\mathbf{S}^{(n)}) \right) \leq \epsilon \|\mathcal{X}\|$$

Compression ratio:

$$C = \prod_{k=1}^{N} I_n \Big/ \left(\prod_{k=1}^{N} R_n + \sum_{k=1}^{N} I_n R_n \right).$$

Simulation of an autoignitive premixture of air and ethanol in Homogeneous Charge Compression Ignition

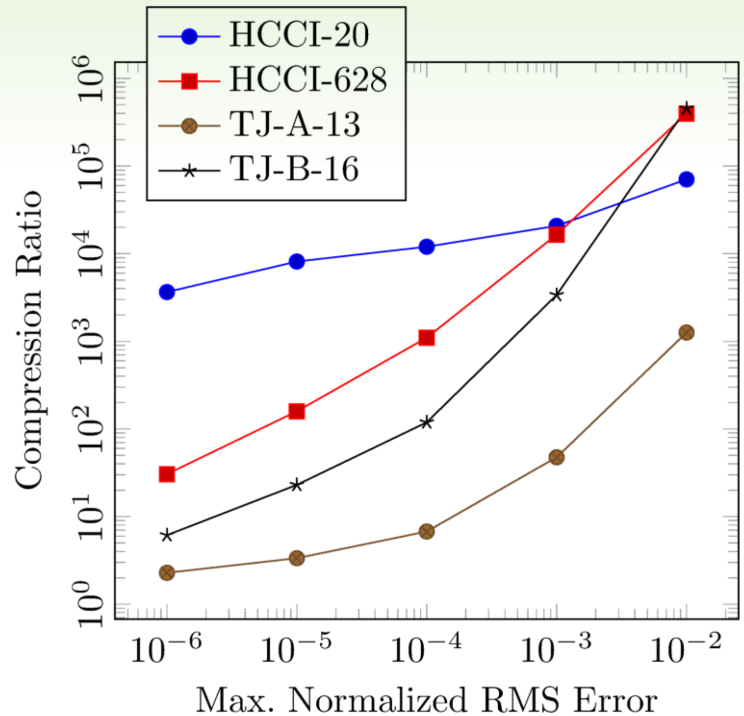
HCCI-628: 672 x 672 x 33 x 628, 72 GB

Temporally-evolving planar slot jet flame with DME (dimethyl ether) as the fuel

TJ-A-13: 300 x 500 x 240 x 35 x 13, 122 GB

TJ-B-16: 460 x 700 x 360 x 35 x 16, 512 GB

Thanks to Hemanth Kolla and Ankit Bhagatwala for combustion application data, from Sandia's S3D direct numerical simulation code



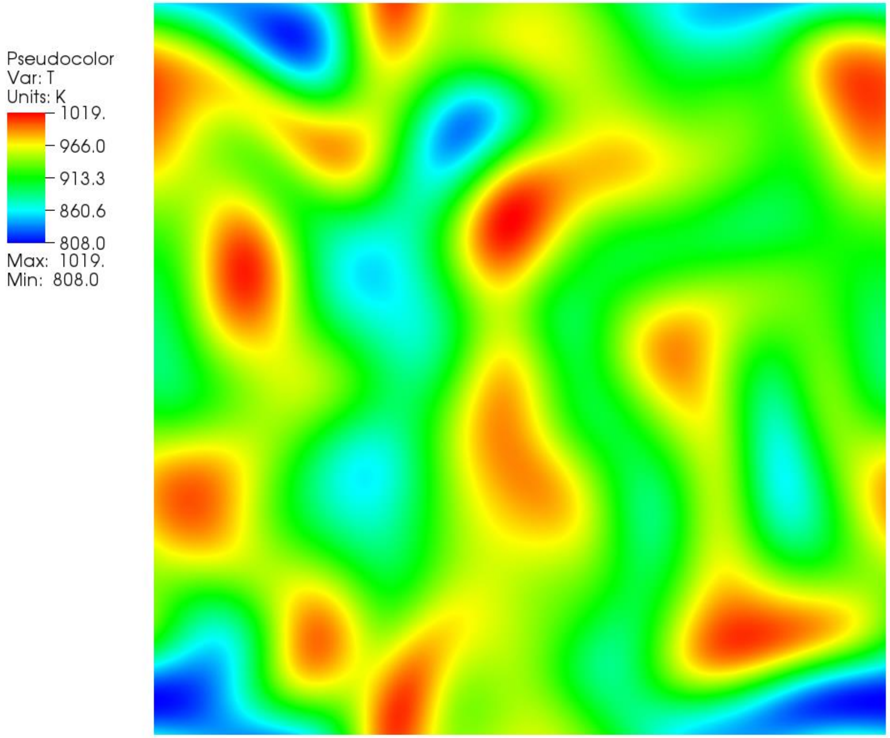
Dataset	Reduced Size	Max. Elem. Error	Comp. Ratio
HCCI-1	(16, 16, 4, 1)	3.6e-5	573
HCCI-20	(20, 18, 6, 5)	2.0e-4	7083
HCCI-628	(192, 183, 16, 104)	1.2e-3	139
TJ-A-1	(257, 139, 186, 20, 1)	1.7e-3	9
TJ-A-13	(300, 209, 240, 25, 13)	3.2e-3	3

Sample Results for one Species in HCCI: Error is Negligible

Compression: 586

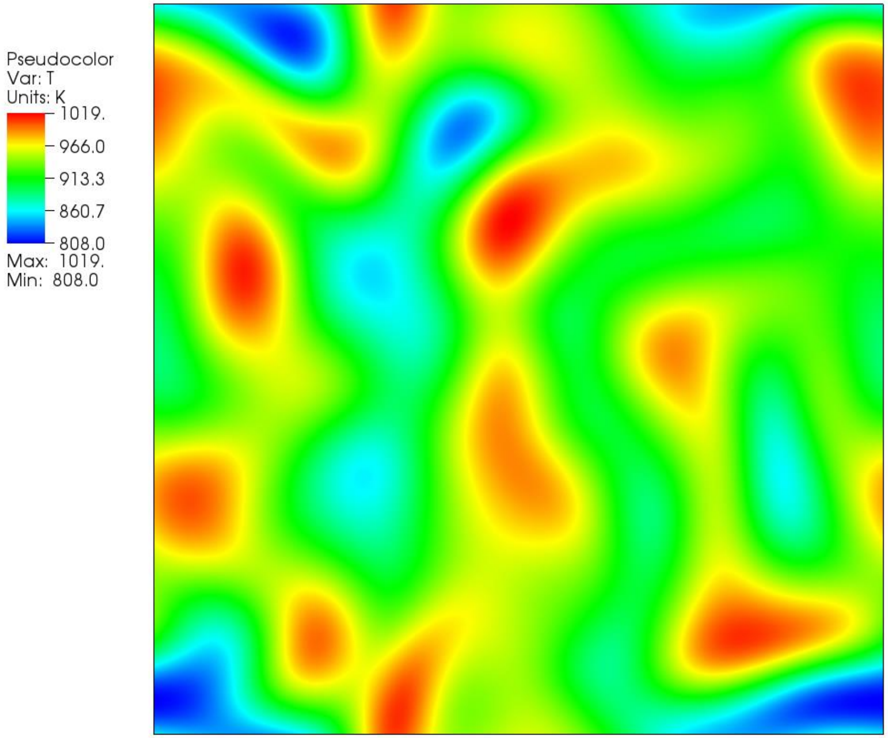
Original $\mathcal{X}$

672 x 672 x 33 x 8



Recovered $\hat{\mathcal{X}}$

48 x 48 x 20 x 3



$\boxed{\text{910MB} \Rightarrow \text{1.5MB}}$

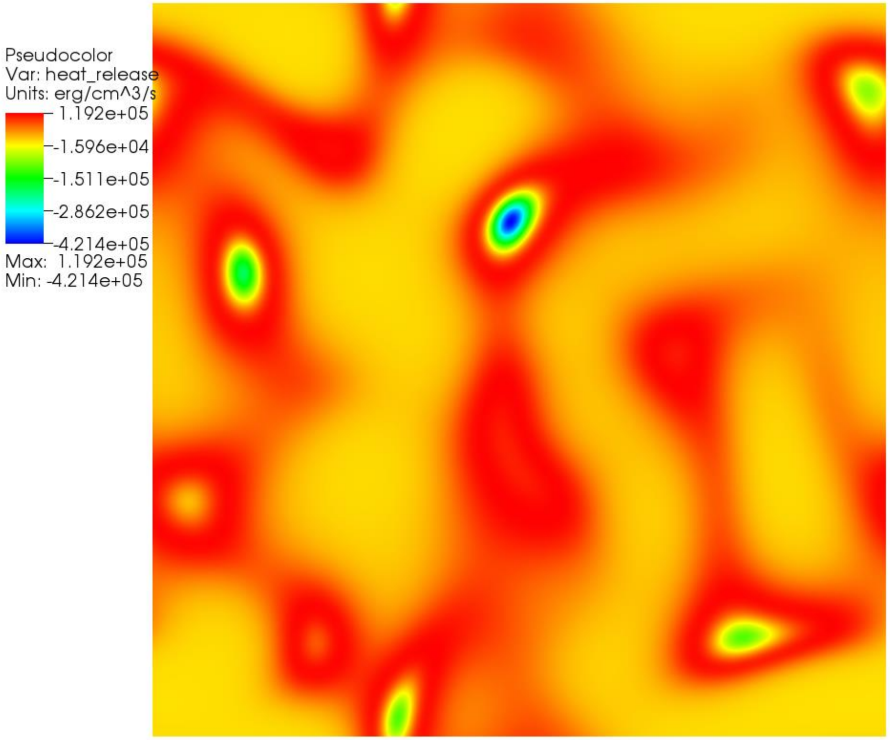
$$\frac{||\mathcal{X} - \hat{\mathcal{X}}||}{||\mathcal{X}||} = 3.08 \times 10^{-8}$$

Sample Results for "Derived" Quantity: Error is Negligible

Compression: 586

Original $\mathcal{X}$

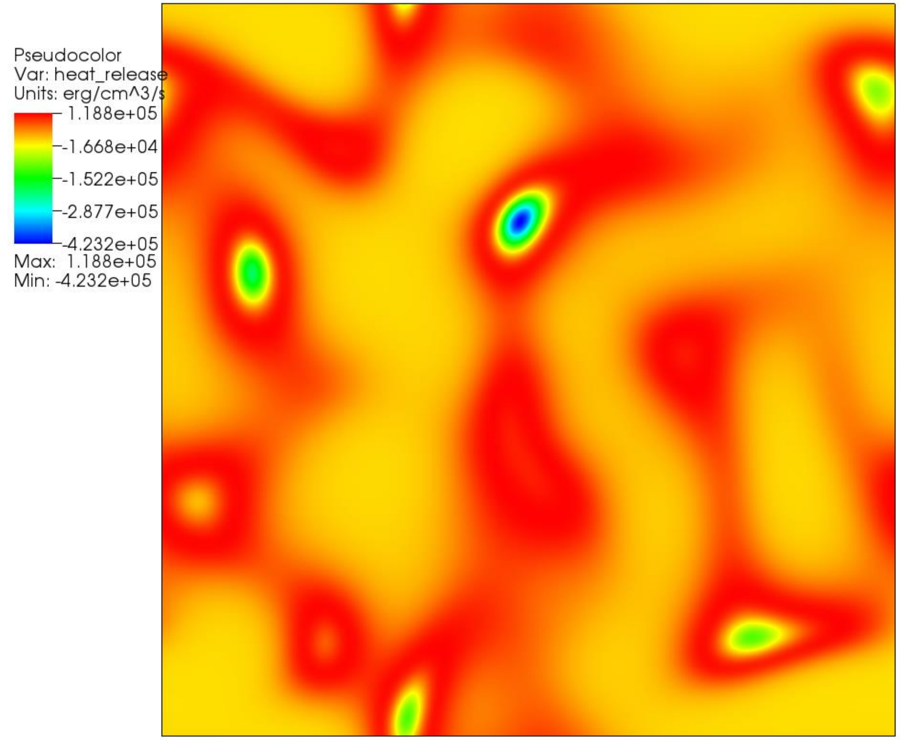
672 x 672 x 33 x 8



$\boxed{\text{910MB} \Rightarrow \text{1.5MB}}$

Recovered $\hat{\mathcal{X}}$

48 x 48 x 20 x 3



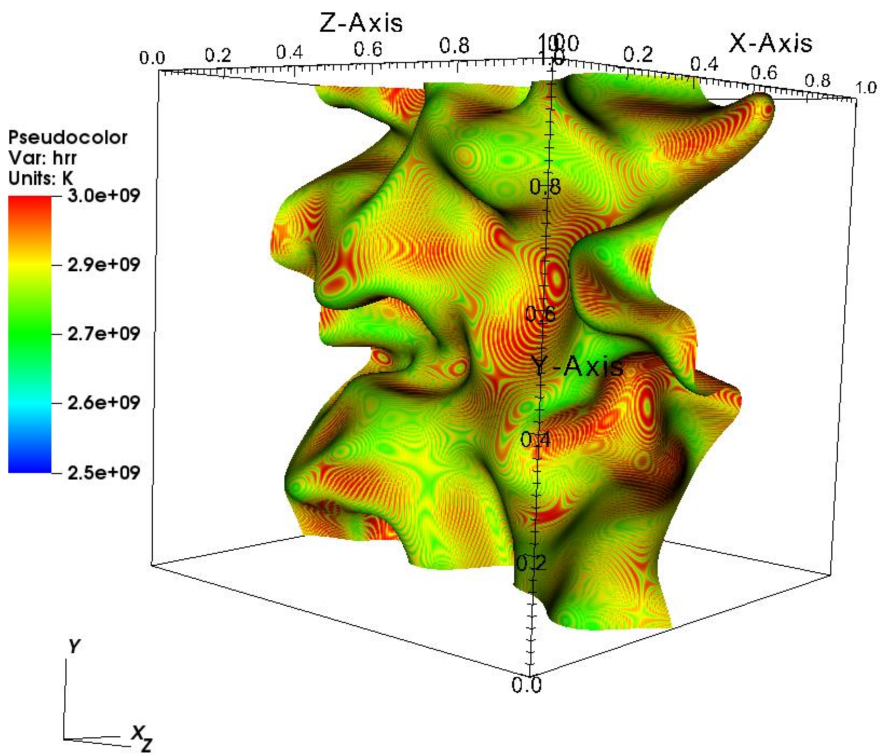
$$\frac{\|\mathcal{X} - \hat{\mathcal{X}}\|}{\|\mathcal{X}\|} = 3.08 \times 10^{-8}$$

Sample Results for one Species in 3D HCCI: Error is Negligible

Compression: 1000

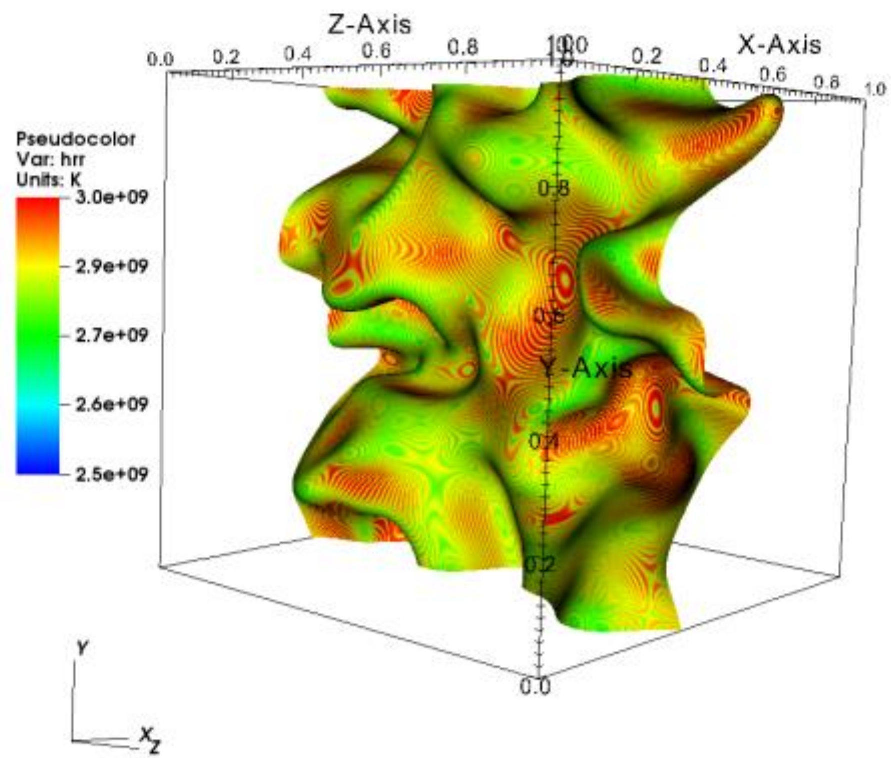
Original x

500 x 500 x 500 x 11 x 5



Recovered $\hat{x}$

50 x 50 x 50 x 11 x 5



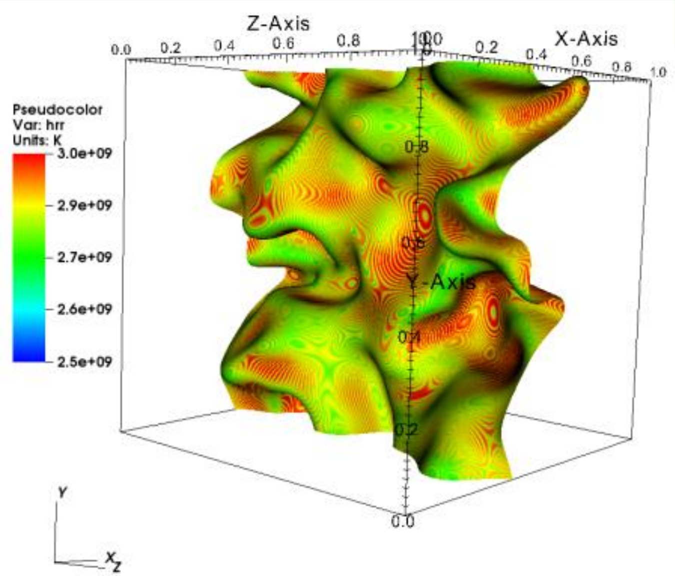
$\boxed{52 \text{ GB} \Rightarrow 52 \text{ MB}}$

Partial Reconstruction

Reconstruction requires as much space as the original data!

$$\hat{\mathcal{X}} = \mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)} \times_4 \mathbf{U}^{(4)} \times_5 \mathbf{U}^{(5)}$$

$I_1 \times I_2 \times I_3 \times I_4 \times I_5$



But we can just reconstruct the portion that we need at the moment:

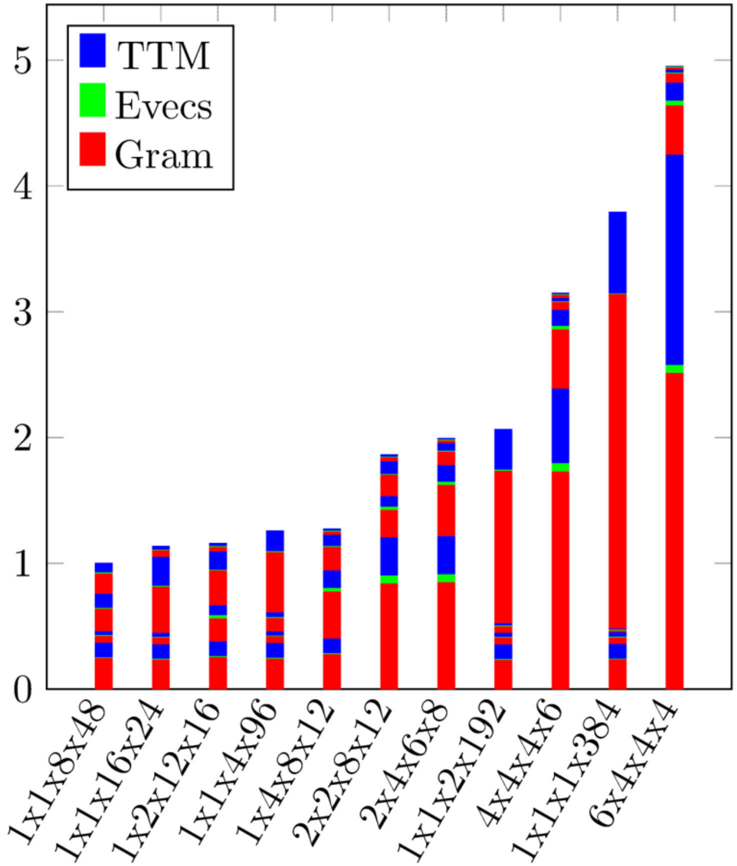
$$\bar{\mathcal{X}} = \mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)} \times_4 \underbrace{\mathbf{U}^{(4)}\mathbf{e}_k}_{\text{Pick out } k\text{th species}} \times_5 \underbrace{\mathbf{U}^{(5)}\mathbf{e}_l}_{\text{Pick out } l\text{th time}}$$

$I_1 \times I_2 \times I_3 \times 1 \times 1$

Parameter Choices: Processor Grid Configuration & Mode Order

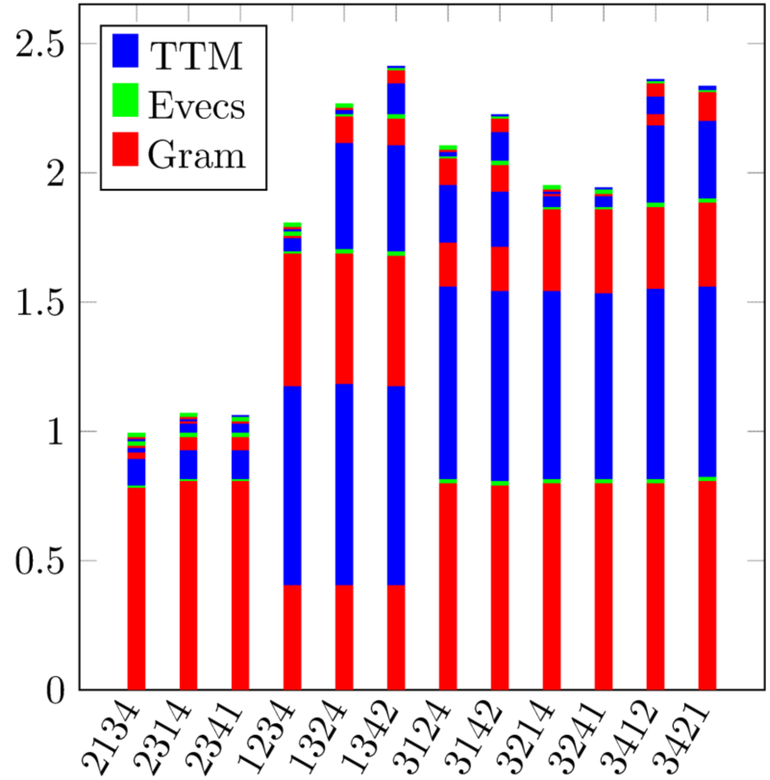
Processor Grid Configuration

$\mathcal{I}$: 384 x 384 x 384 x 384
$\mathcal{R}$: 96 x 96 x 96 x 96



Mode Order

$\mathcal{I}$: 25 x 250 x 250 x 250
$\mathcal{R}$: 10 x 10 x 100 x 100



Strong & Weak Scaling

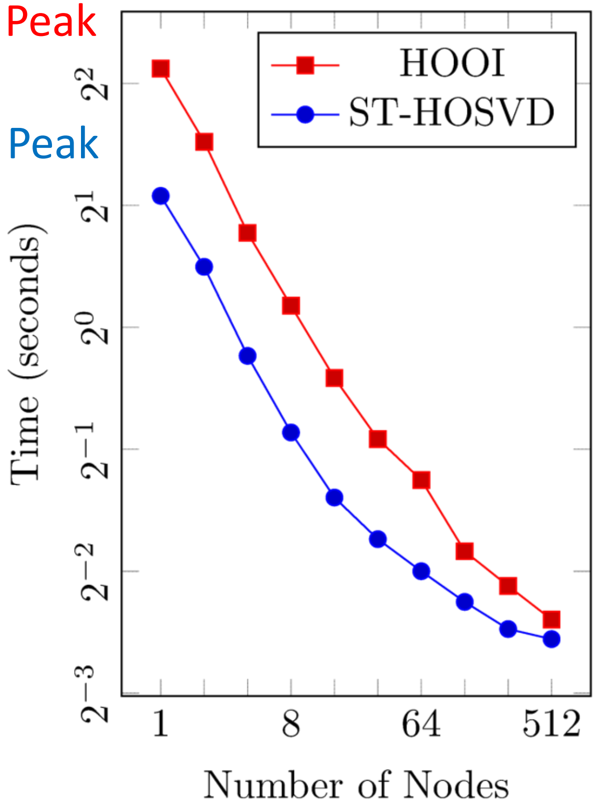
Strong Scaling with 24 x 2^k processors

I: 500 x 300 x 240 x 35

R: 42 x 115 x 81 x 19

40% Peak

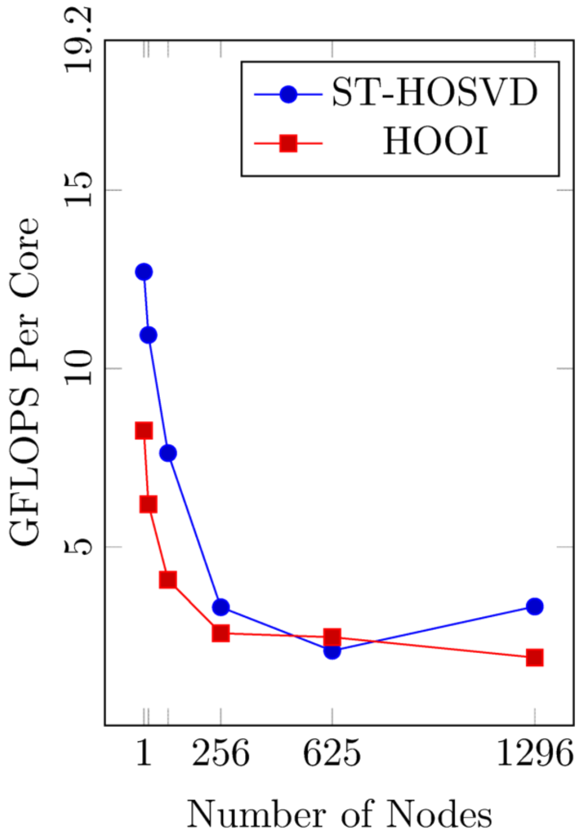
83% Peak



Weak Scaling with 24 x k^4 processors

I: $200k$ x $200k$ x $200k$ x $200k$

R: $20k$ x $20k$ x $20k$ x $20k$

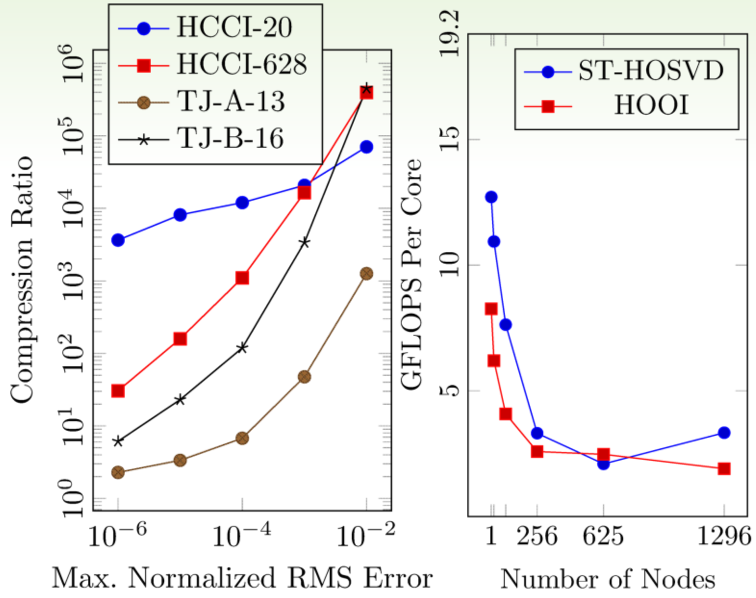


70 secs

15 TB

Parallel Tucker Compression

- First-ever implementation of distributed-memory parallel Tucker decomposition
  - Avoids unnecessary data permutations
- Up to 10^5 compression on real-world data with minimal loss in accuracy
- Scales well – achieving 17% of peak on 1000s of processors
- Future work
  - Detailed application studies
  - Use QR instead of Gram



For more information:
Tammy Kolda,
tgkolda@sandia.gov

W. Austin, G. Ballard, and T. G. Kolda, ***Parallel Tensor Compression for Large-Scale Scientific Data***, arXiv, October 2015, submitted for publication

Backup Slides

Sandia National Laboratories

Acknowledgements

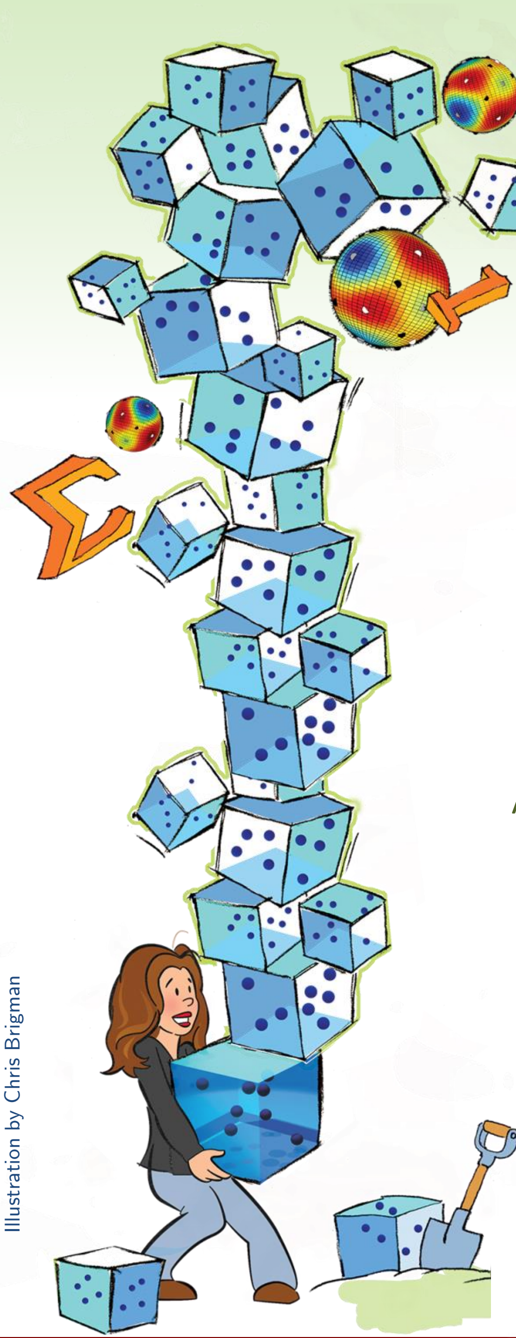
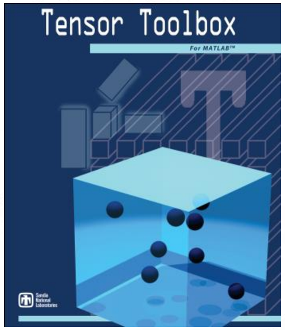


Illustration by Chris Brigman

Kolda and Bader, Tensor Decompositions and Applications, *SIAM Review*, 2009



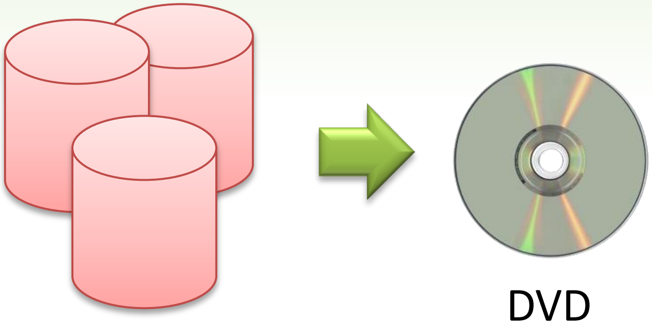
Tensor Toolbox for MATLAB
Bader, Kolda, Acar, Dunlavy, and others

<u>Co-authors</u>

- Evrim Acar (Univ. Copenhagen*)
- Woody Austin (Univ. Texas Austin*)
- Brett Bader (Digital Globe*)
- Grey Ballard (Sandia)
- Eric Chi (NC State Univ.*)
- Danny Dunlavy (Sandia)
- Sammy Hansen (IBM*)
- Joe Kenny (Sandia)
- Jackson Mayo (Sandia)
- Morten Mørup (Denmark Tech. Univ.)
- Todd Plantenga (FireEye*)
- Martin Schatz (Univ. Texas Austin*)
- Teresa Selee (GA Tech Research Inst.*)
- Jimeng Sun (GA Tech)

Plus many more collaborators for workshops, tutorials, etc.

** = Worked for Sandia at some point*



DVD

Parallel Filesystem