Parallel Tensor Compression for Large-Scale Scientific Data

Woody Austin
Univ. Texas, Austin, TX
Grey Ballard* and Tamara G. Kolda
Sandia National Laboratories, Livermore, CA

DMML Workshop
Berkeley CA
October 23, 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

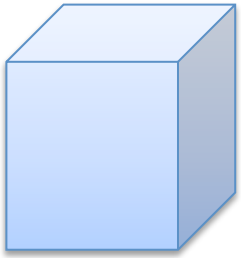
Vector
N = 1

$\mathbf{x}$

Matrix
N = 2

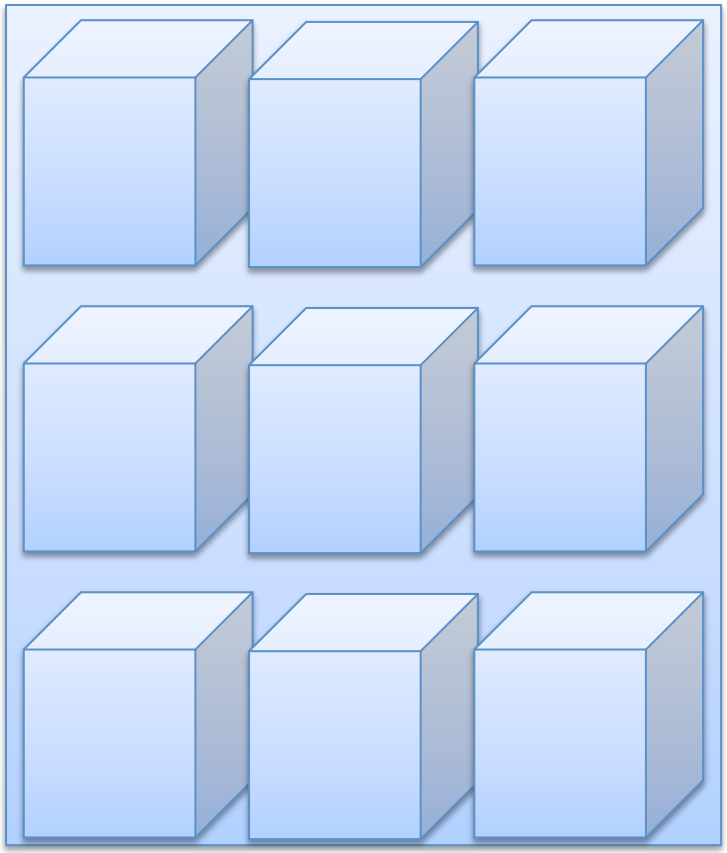
$\mathbf{X}$

3rd-Order Tensor
N = 3



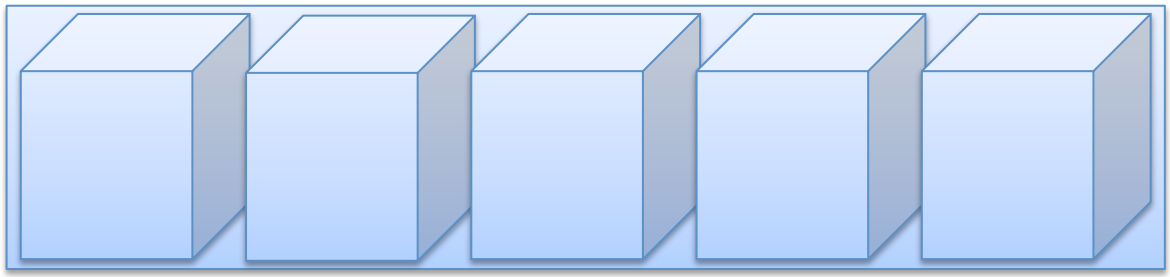
$\mathcal{X}$

5th-Order Tensor
N = 5



$\mathcal{X}$

4th-Order Tensor
N = 4

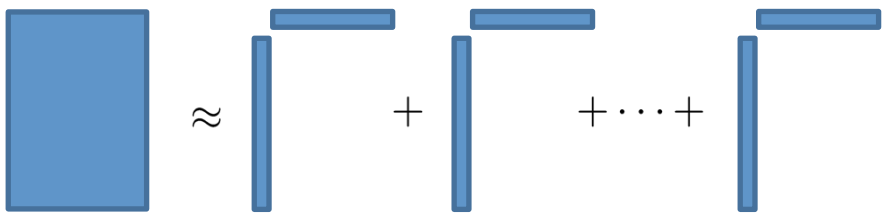


$\mathcal{X}$

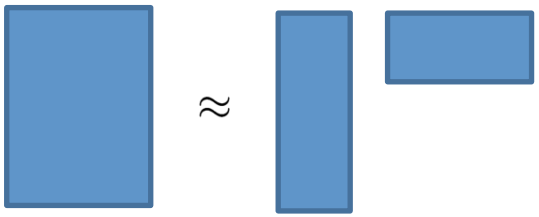
Tensor decompositions are the new matrix decompositions

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

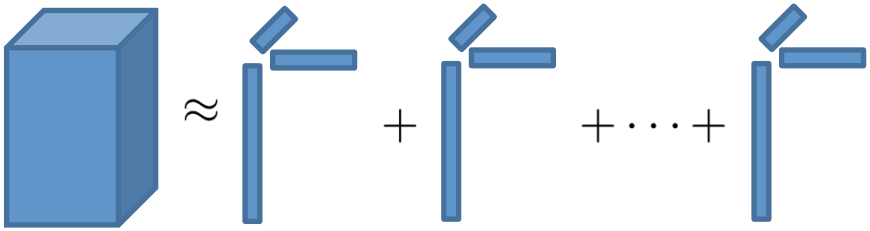
Viewpoint 1: Sum of outer products, useful for interpretation



Viewpoint 2: High-variance subspaces, useful for compression

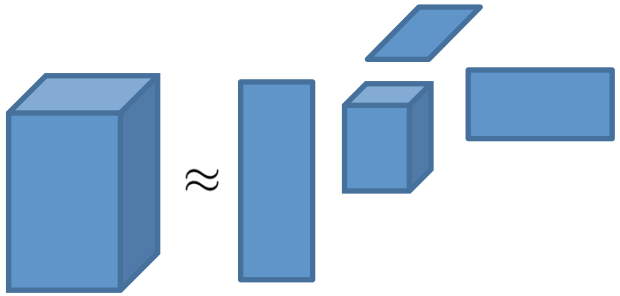


CP Model: Sum of d-way outer products, useful for interpretation



CANDECOMP, PARAFAC, Canonical Polyadic, CP

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

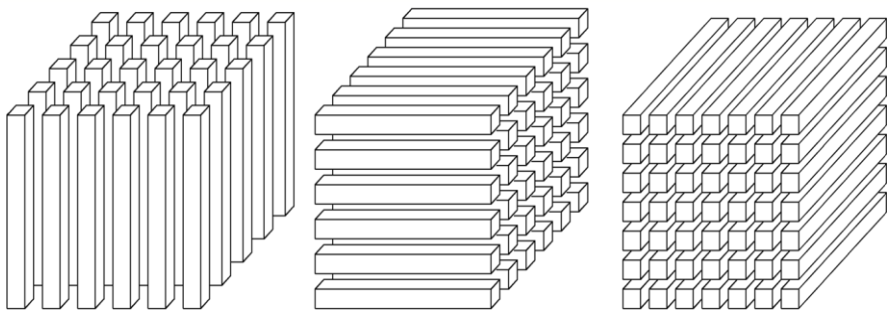


HOSVD, Best Rank-(R1,R2,…,RN) decomposition

Other models for compression include hierarchical Tucker and tensor train.

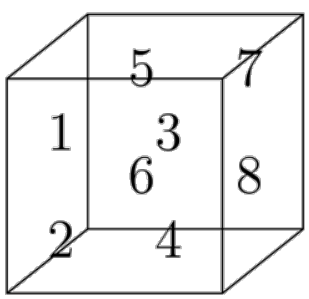
Tensor fibers, mode-n unfolding, and mode-n Multiplication

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



Mode-1 Fibers Mode-2 Fibers Mode-3 Fibers

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

$\mathcal{X} =$ 

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

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

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

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

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

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

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

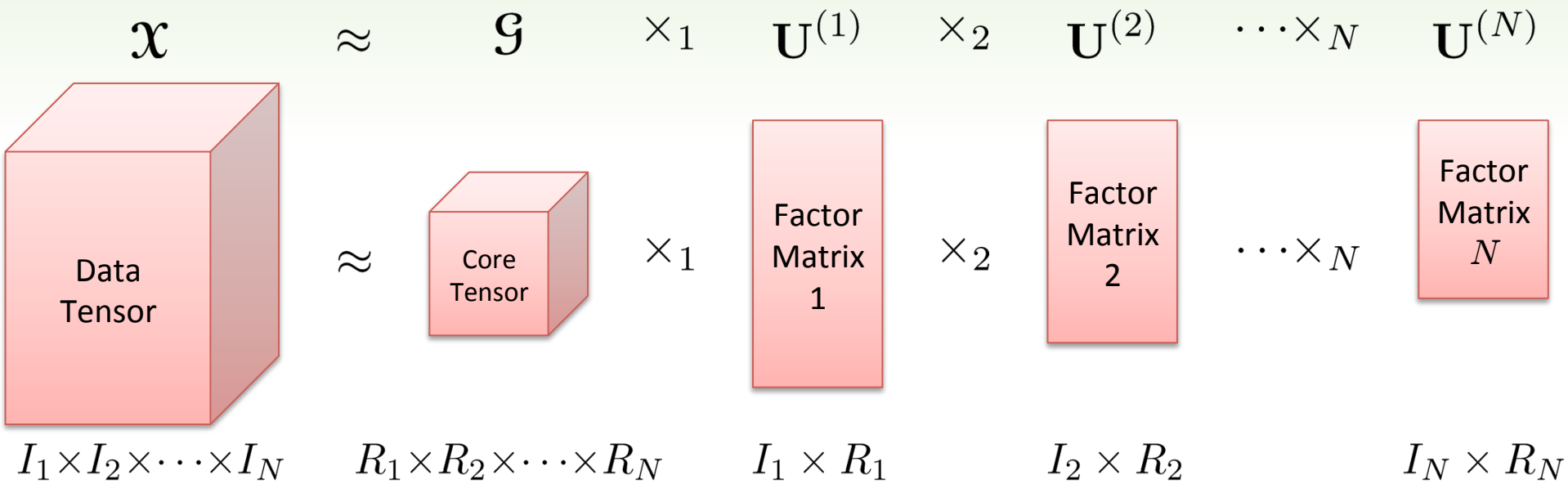
Equivalent to matrix operation:

$$K \times I_n \qquad I_n \times \hat{I}_n$$

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

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

Tucker decomposition



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

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>Couple Facts</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

$$\max_{\{\mathbf{U}^{(n)}\}} \sum_{i_1 \ldots i_N} g^2_{i_1 \ldots i_N} \text{ subject to } \mathcal{G} = \mathcal{X} \times \{ \mathbf{U}^{(n)\mathsf{T}} \}$$

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

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

Truncated Higher-Order SVD (HOSVD) is a sequence of truncated SVDs

```
procedure T-HOSVD(𝒳, {R_n})
    for n = 1, ..., N do
        U^(n) ← leading R_n left singular vectors of X_(n)
    end for
    𝒢 ← 𝒳 × {U^(n)T}
    return (𝒢, { U^(n) })
end procedure
```

Also known as "Tucker1" method.

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

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

Truncated HOSVD is a sequence of truncated SVDs

```
procedure T-HOSVD(𝒳, {R_n})
    for n = 1, ..., N do
        S^(n) = X_(n) X_(n)^T
        U^(n) ← leading R_n eigenvectors of S^(n)
    end for
    𝒢 ← 𝒳 × {U^(n)⊤}
    return (𝒢, { U^(n) })
end procedure
```

Also known as "Tucker1" method.

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

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

Sequentially Truncated HOSVD improves further

procedure ST-HOSVD($\mathcal{X}$, $\{R_n\}$)
  $\mathcal{Y} \leftarrow \mathcal{X}$
  for $n = 1, \ldots, N$ **do**
    $\mathbf{S}^{(n)} \leftarrow \mathbf{Y}_{(n)}\mathbf{Y}_{(n)}^{\mathsf{T}}$
    $\mathbf{U}^{(n)} \leftarrow$ 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

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

Vannieuwenhoven, Vandebril, and Meerbergen (2012)

Higher-Order Orthogonal Iteration (HOOI) improves again

Sandia National Laboratories

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

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

Key kernels are TTM and Gram

procedure ST-HOSVD($\mathcal{X}$, $\{R_n\}$)

$\mathcal{Y} \leftarrow \mathcal{X}$

for $n = 1, \ldots, N$ **do**

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

$\mathbf{U}^{(n)} \leftarrow$ leading R_n eigenvectors of $\mathbf{S}^{(n)}$

TTM $\mathcal{Y} \leftarrow \mathcal{Y} \times_n \mathbf{U}^{(n)\mathsf{T}}$

end for

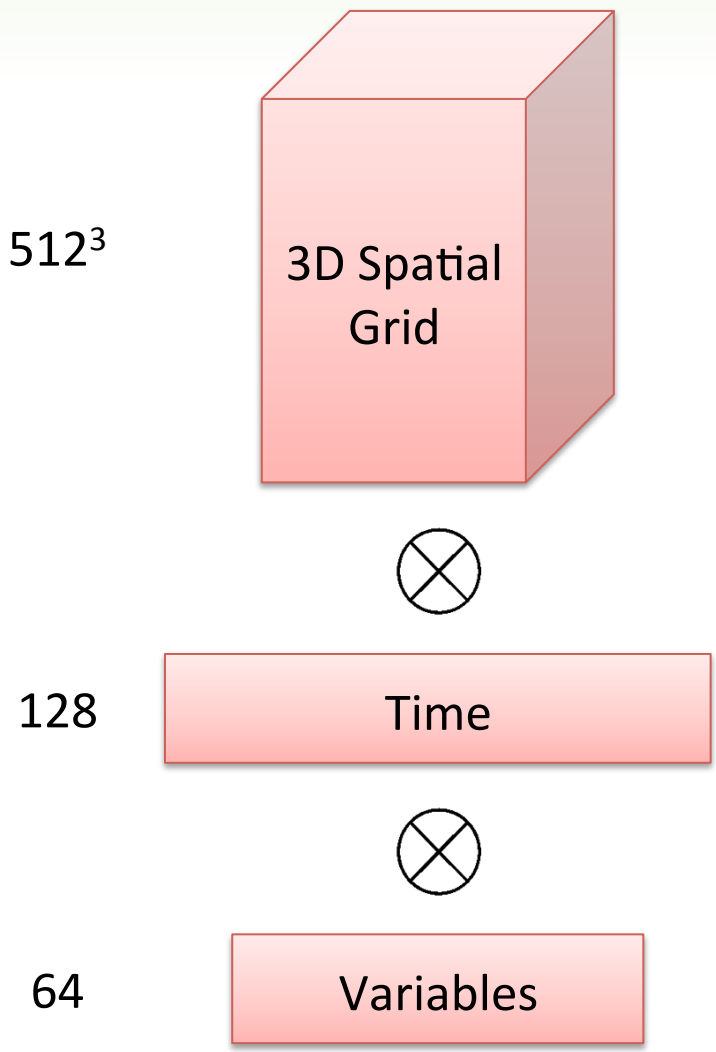
$\mathcal{G} \leftarrow \mathcal{Y}$

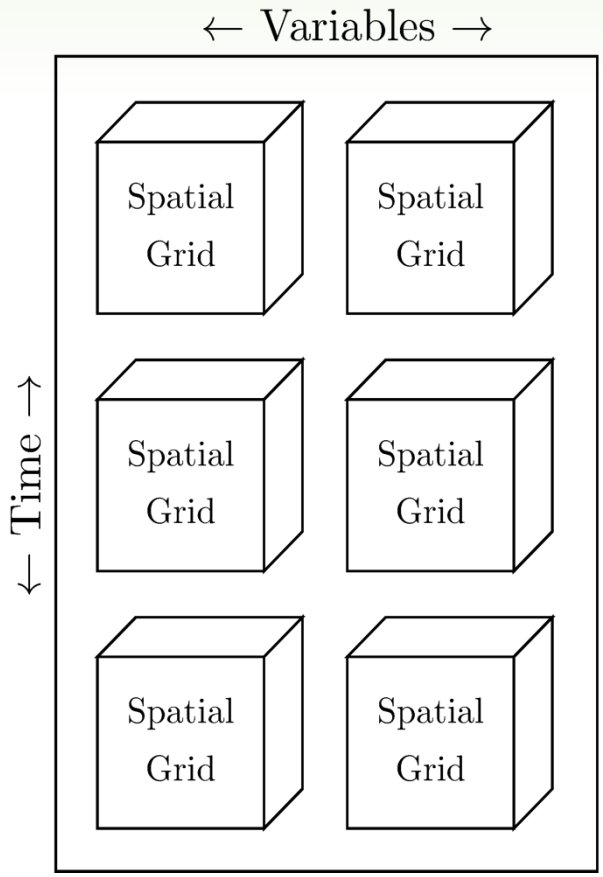
return $(\mathcal{G}, \{\mathbf{U}^{(n)}\})$

end procedure

Vannieuwenhoven, Vandebril, and Meergergen (2012)

Tensors in scientific applications are huge, need parallel methods

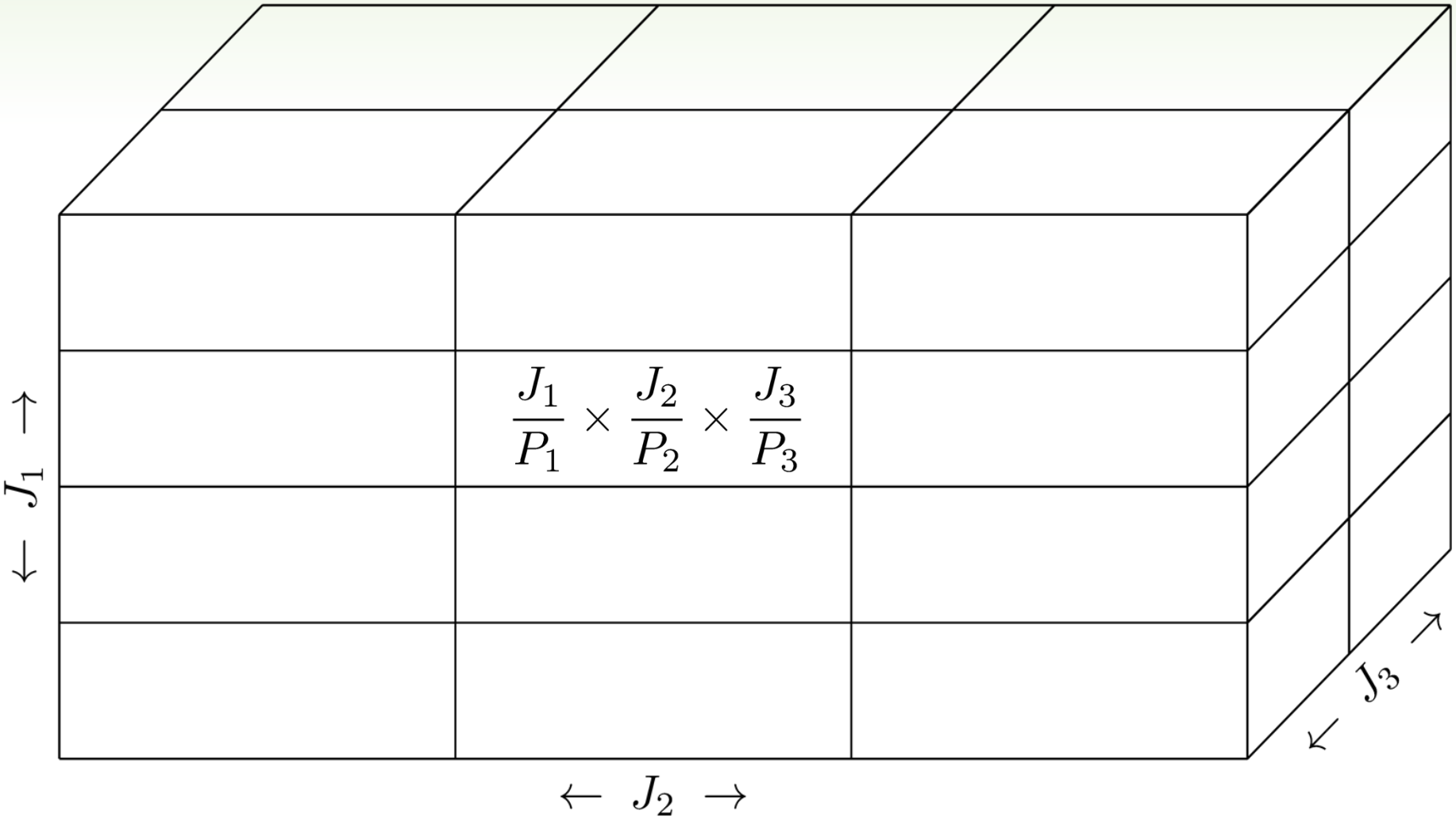




2^{40} elements

8 TB (double precision)

Tensor distribution: Cartesian



Processor Grid: $P_1 \times P_2 \times P_3 = 4 \times 3 \times 2$

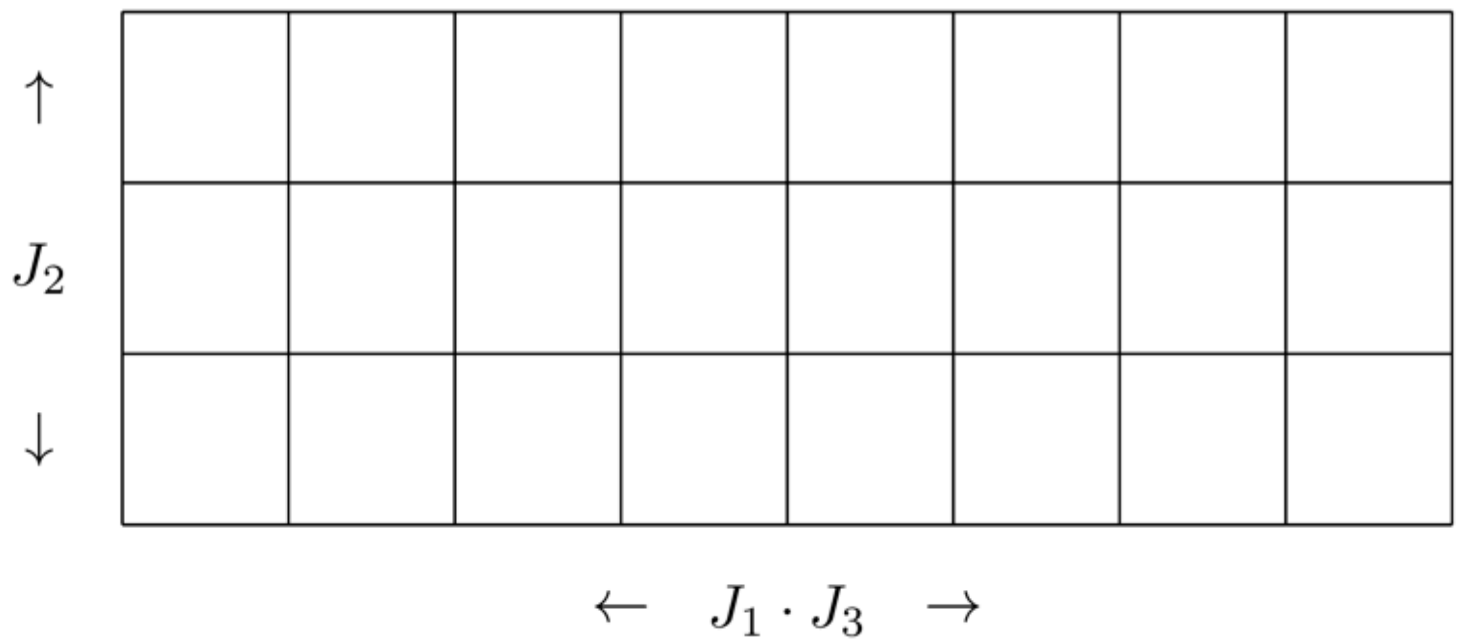
Unfolded tensor distribution

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

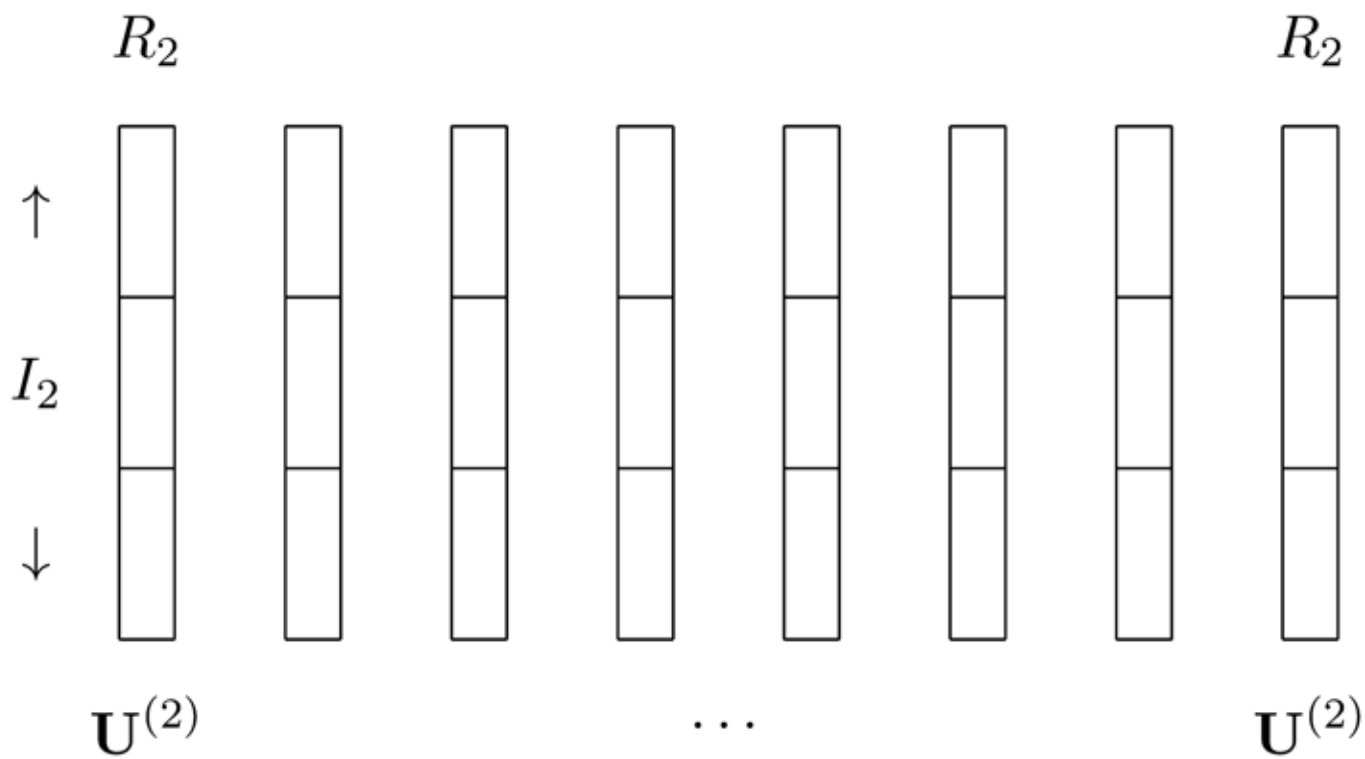
Global Unfolded Tensor: $J_n \times \hat{J}_n$

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



Redundant factor matrix distribution

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



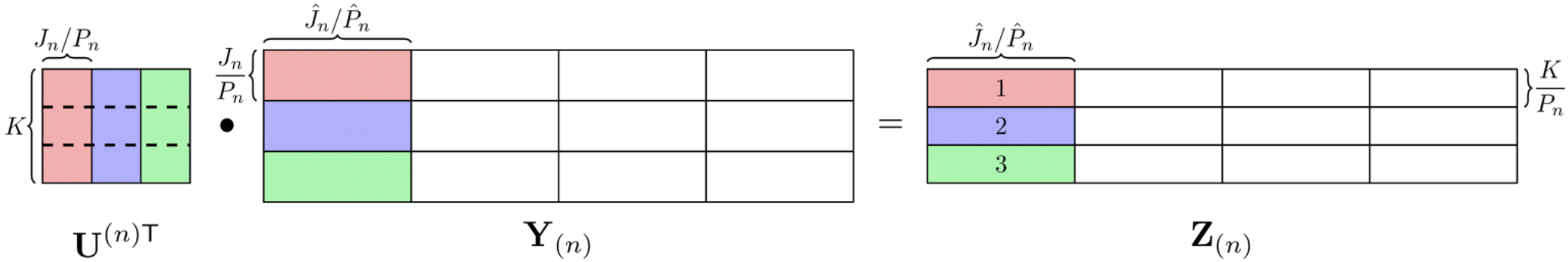
Processor Grid: $4 \times 3 \times 2$

Parallel TTM

$$\mathcal{Z} = \mathcal{Y} \times_n \mathbf{U}^{(n)\mathsf{T}} \quad \Leftrightarrow \quad \mathbf{Z}_{(n)} = \mathbf{U}^{(n)\mathsf{T}} \mathbf{Y}_{(n)}$$

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

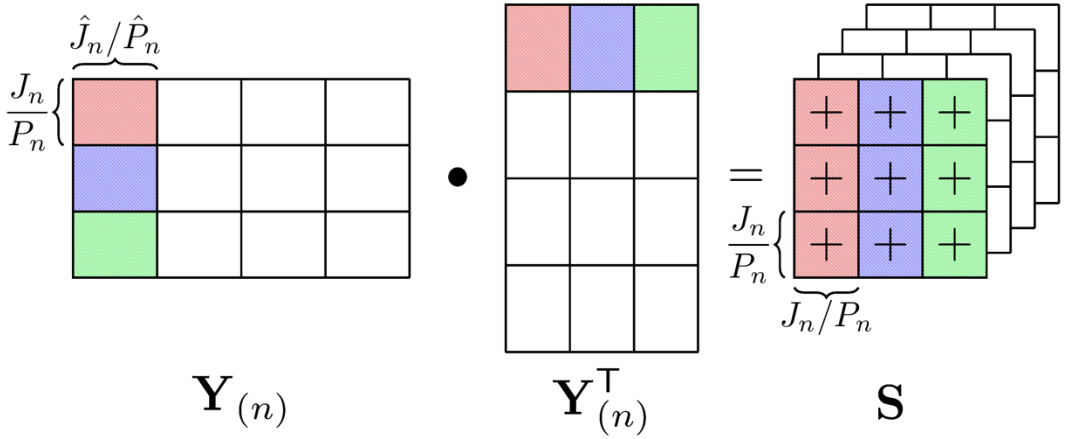
$$C_{\text{TTM}} = 2\gamma \frac{JK}{P} + \alpha P_n \log P_n + \beta (P_n - 1) \frac{\hat{J}_n K}{P}$$

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

Parallel Gram

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

procedure GRAM($\mathcal{Y}$,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 $\bar{\mathcal{Y}}$ to process $(p_1, \ldots, p_{n-1}, j, \ldots, p_N)$
    Receive $\mathcal{W}$ 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



$$C_{\text{GRAM}} = \gamma 2 J_n J/P + 2(P_n - 1)(\alpha + \beta J/P) + 2\alpha \log \hat{P}_n + 2\beta(\hat{P}_n - 1)J_n^2/P$$

$$M_{\text{GRAM}} = \underbrace{J/P}_{\bar{\mathcal{Y}}} + \underbrace{J/P}_{\mathcal{W}} + \underbrace{J_n^2/P_n}_{\mathbf{V}} + \underbrace{J_n^2/P_n}_{\bar{\mathbf{S}}}$$

Parallel eigenvector computation

```
procedure EIGENVECTORS(S̄, R_n, n)
    myProcID = (p_1, p_2, ..., p_N)
    myProcCol ← (p_1, ..., p_{n-1}, *, p_{n+1}, ..., p_N)
    S = ALL-GATHER(S̄, myProcCol)
    U^(n) = LOCAL-EIGENVECTORS(S, R_n)
    Ū^(n) = ROW-SUBSET(U^(n), P_n, p_n)          ▷ Extract p_n-th block
    return Ū^(n)
end procedure
```

Every processor redundantly computes the leading eigenvectors of the Gram matrix

Application results: compression versus accuracy

Ranks depend on error:

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

Compression ratio:

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

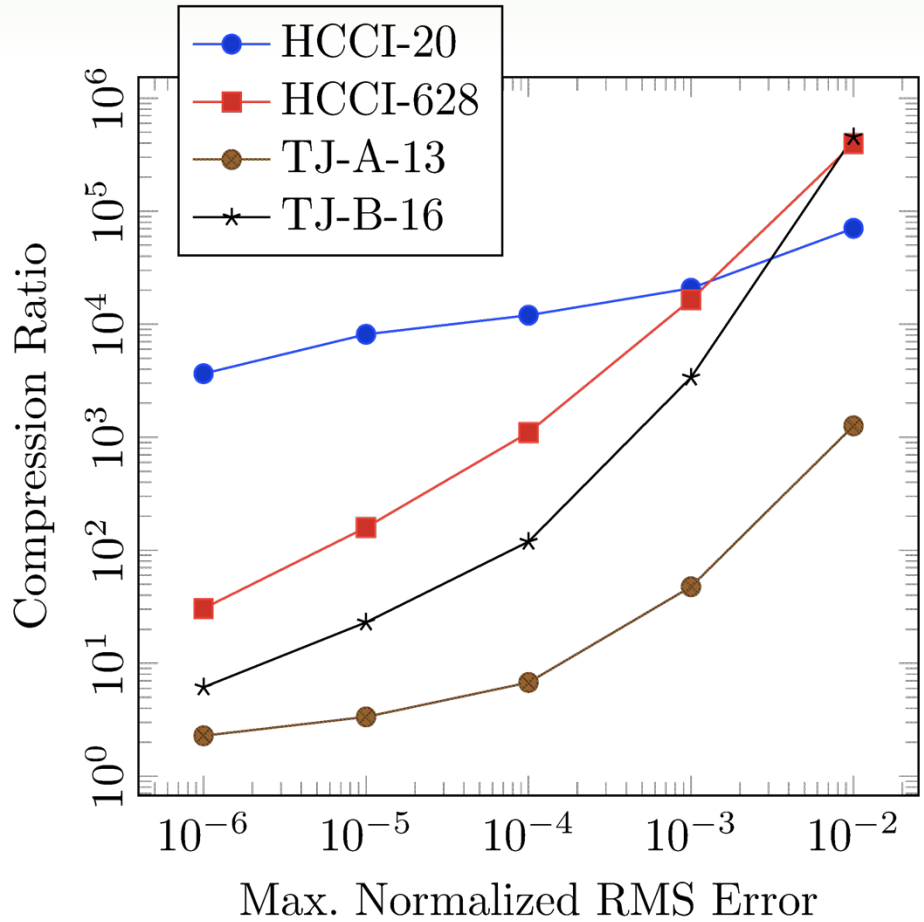
Simulation of an autoignitive premixture of air and ethanol in Homogeneous Charge Compression Ignition

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



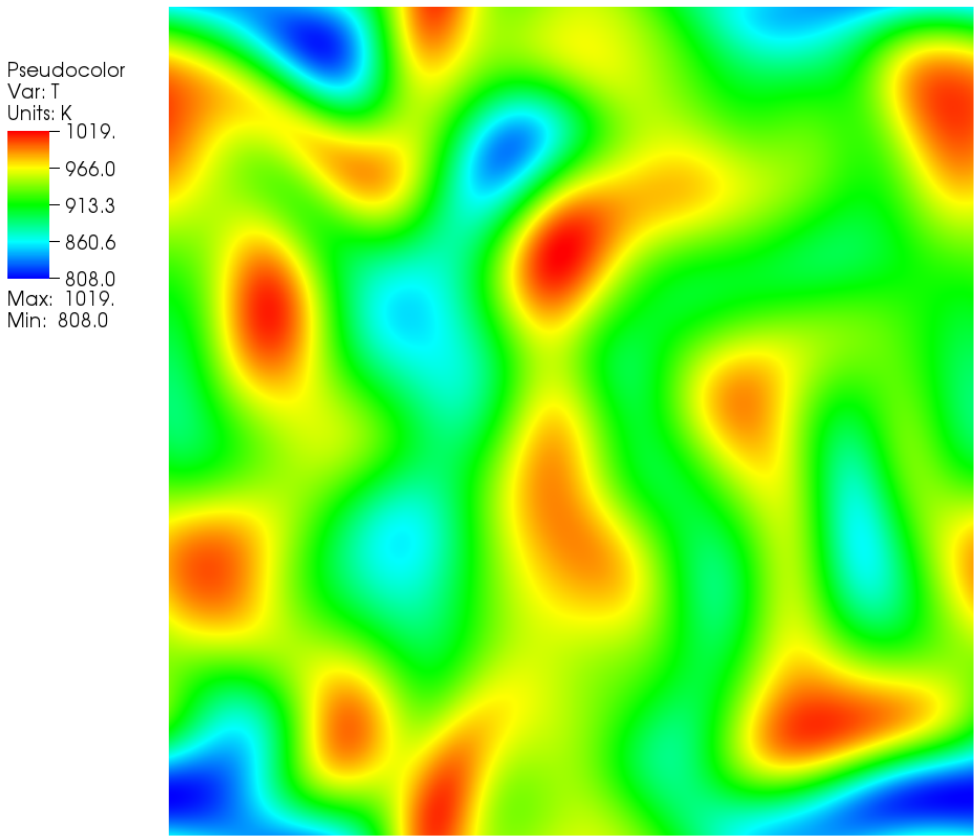
Thanks to Hemanth Kolla and Ankit Bhagatwala for combustion application data, from Sandia's S3D direct numerical simulation code

Sample results for one species in HCCI: error is negligible

Compression: 586

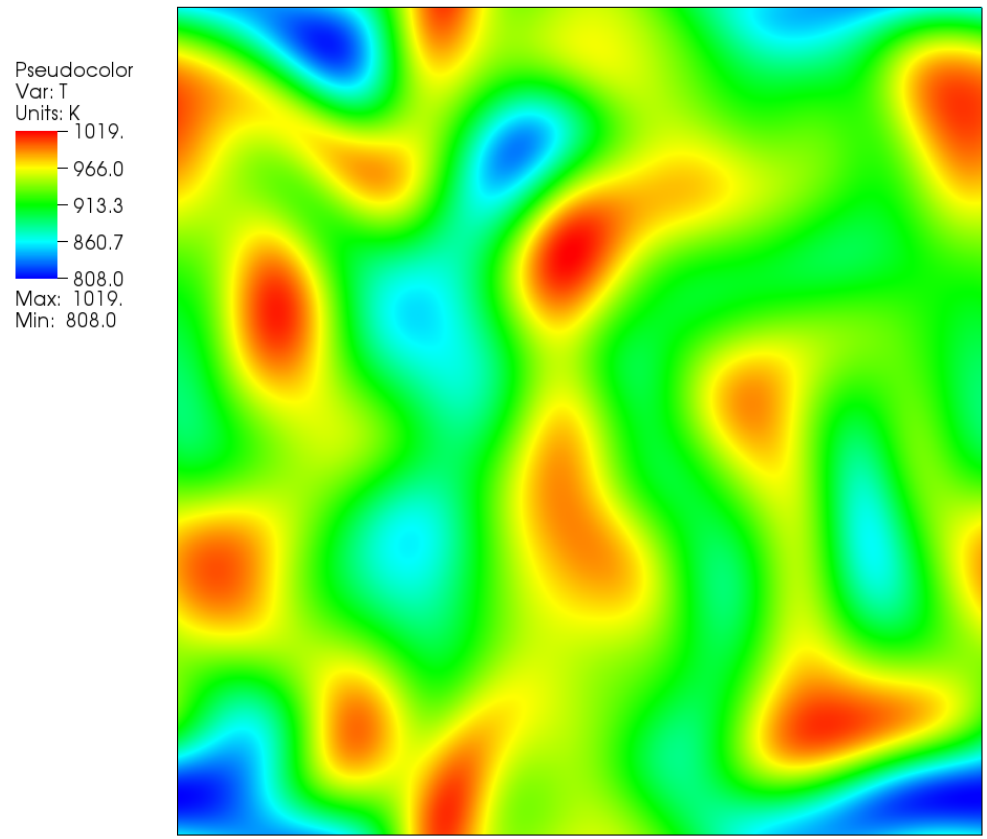
Original $\mathcal{X}$

672 x 672 x 33 x 8



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

48 x 48 x 20 x 3



910MB compressed to 1.5MB

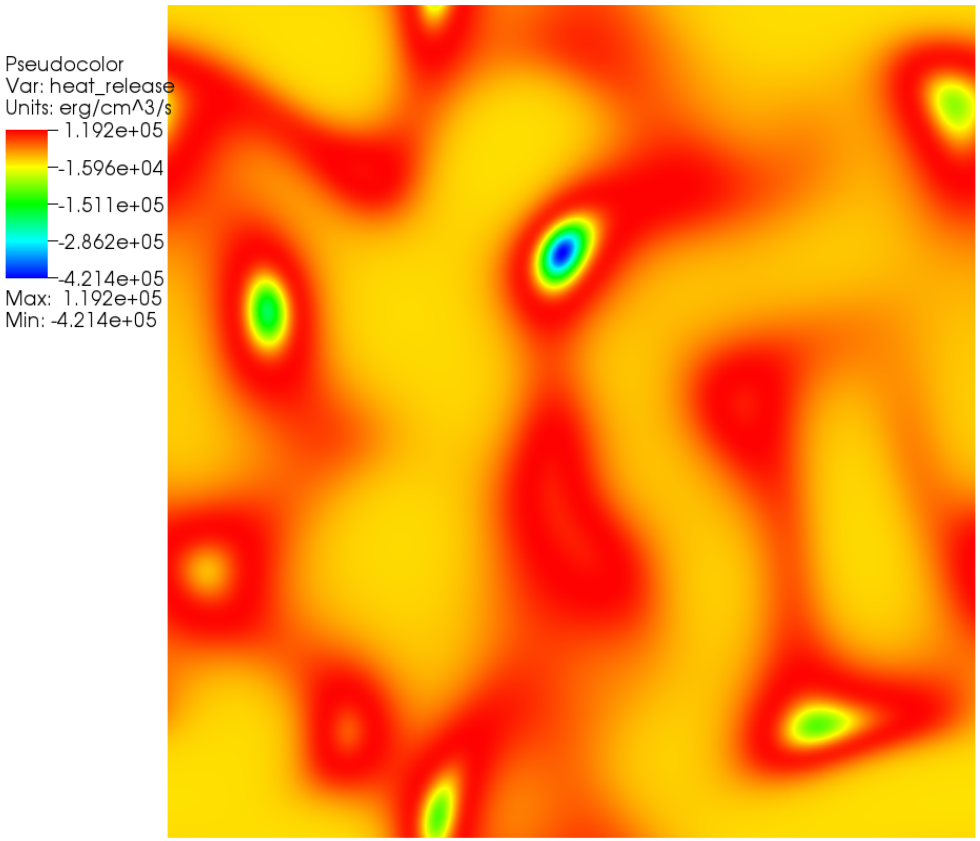
$$\frac{||\mathcal{X} - \hat{\mathcal{X}}||}{||\mathcal{X}||} = 3.08 \times 10^{-8}$$

Sample results for "derived" quantity: error is negligible

Compression: 586

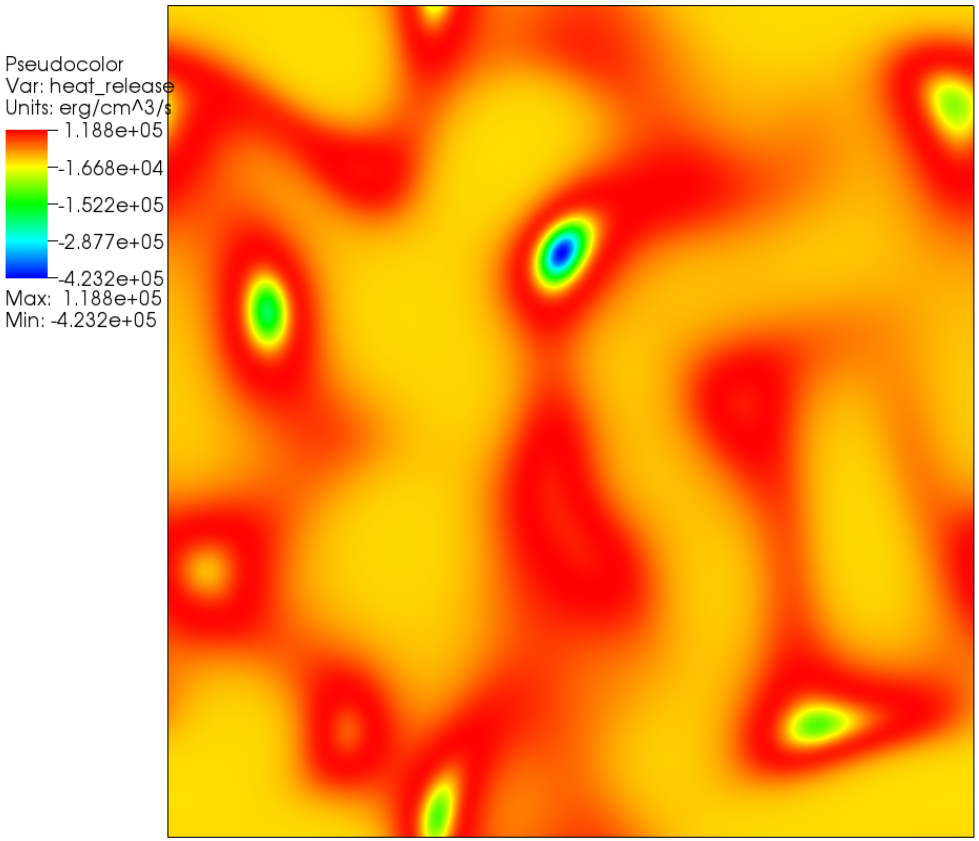
Original $\mathcal{X}$

672 x 672 x 33 x 8



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

48 x 48 x 20 x 3



910MB compressed to 1.5MB

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

Sample results for one species in 3D HCCI: error is negligible

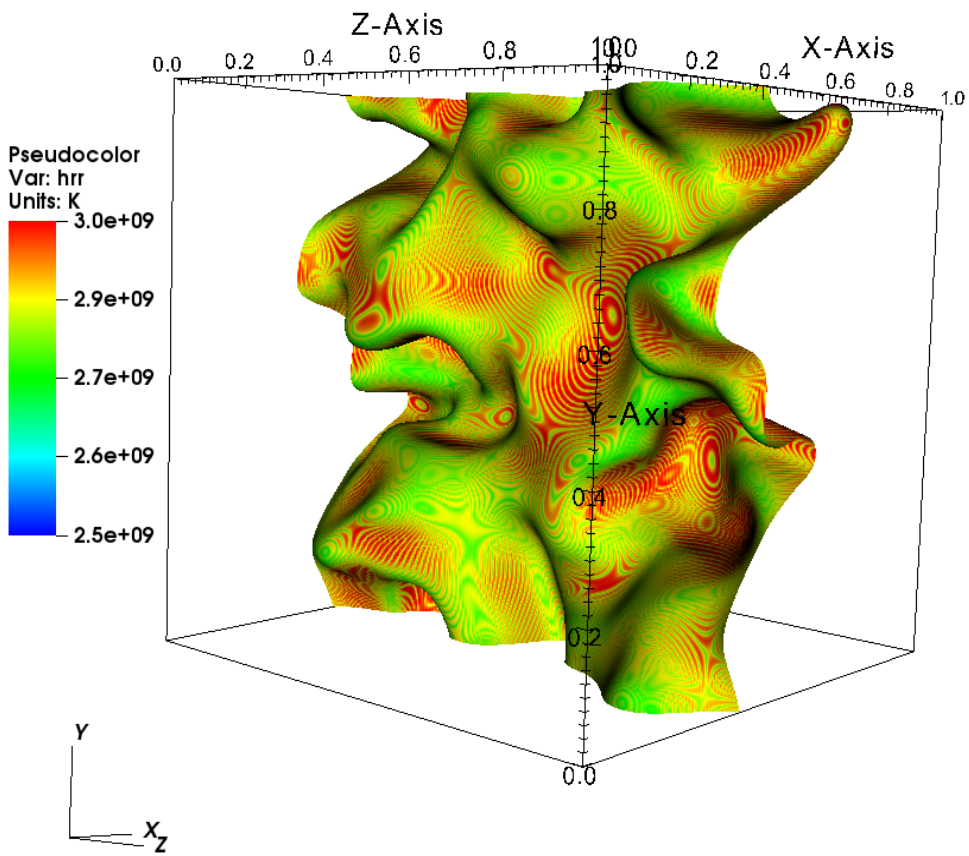
Original $\mathcal{X}$

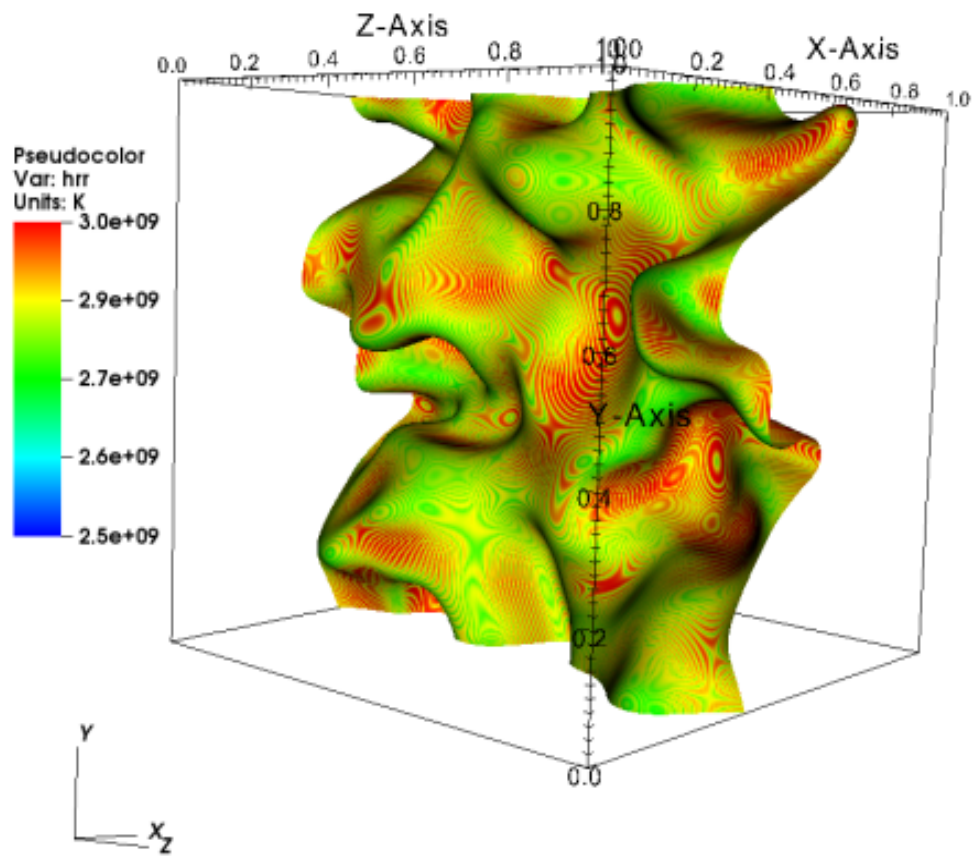
500 x 500 x 500 x 11 x 5

Compression: 1000

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

50 x 50 x 50 x 11 x 5





52 GB compressed to 52 MB

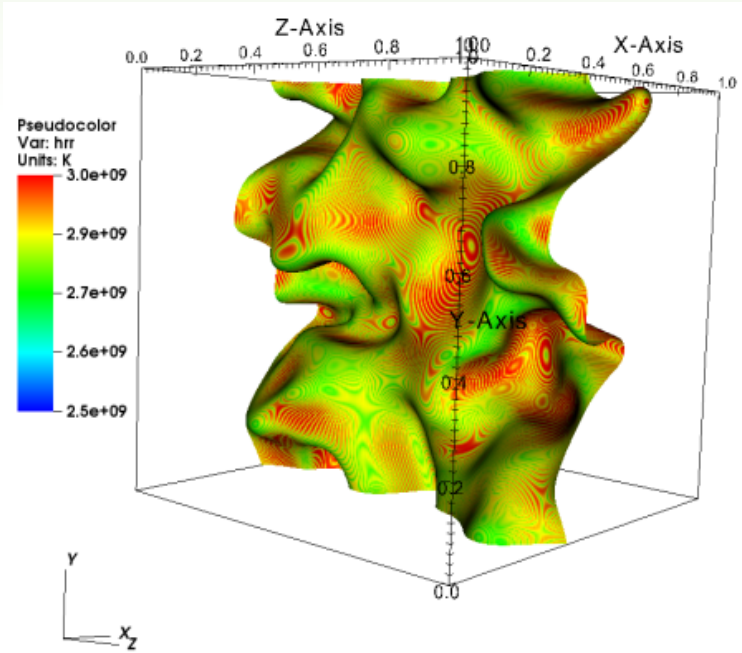
$$\frac{\|\mathcal{X} - \hat{\mathcal{X}}\|}{\|\mathcal{X}\|} = 3.3 \times 10^{-9}$$

Partial reconstruction

Reconstruction requires as much space as the original data!

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

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



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

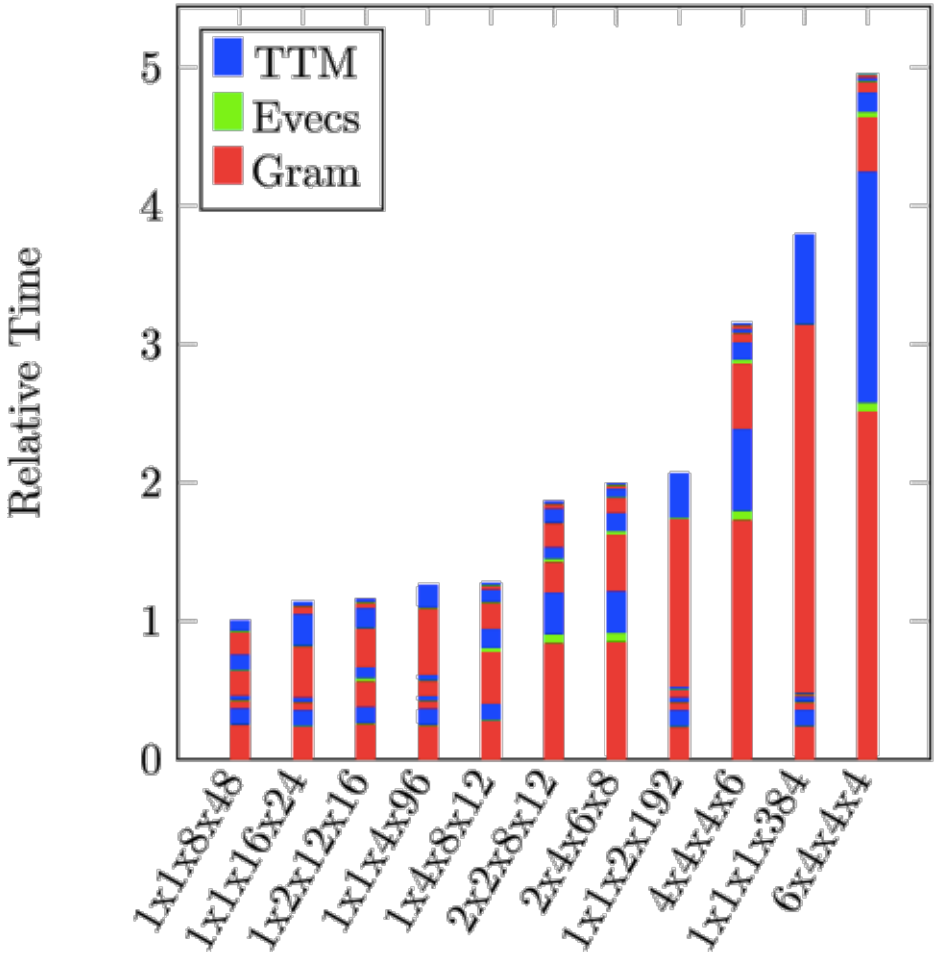
$$\bar{\mathfrak{X}} = \mathfrak{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)} \times_4 \underbrace{\mathbf{U}^{(4)}\mathbf{e}_k}_{\text{Pick out } k\text{th species}} \times_5 \underbrace{\mathbf{U}^{(5)}\mathbf{e}_l}_{\text{Pick out } l\text{th time step}}$$

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

Parameter choices: processor grid configuration & mode order

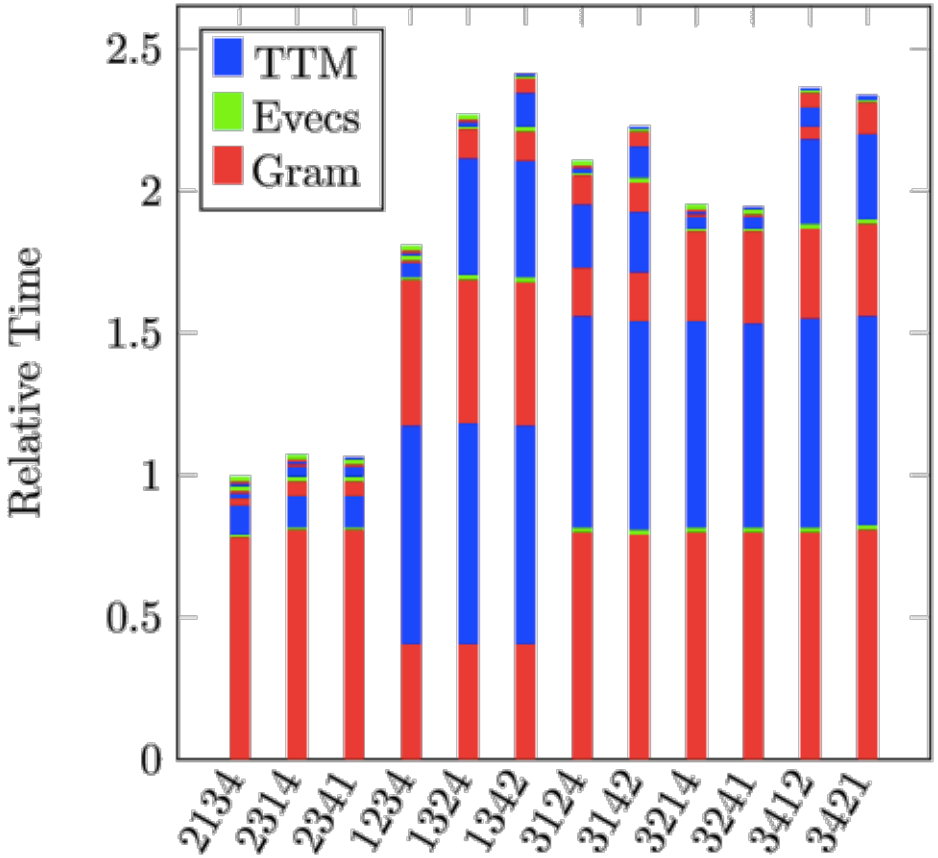
Processor Grid Configuration

I: 384 x 384 x 384 x 384
R: 96 x 96 x 96 x 96



Mode Order

I: 25 x 250 x 250 x 250
R: 10 x 10 x 100 x 100

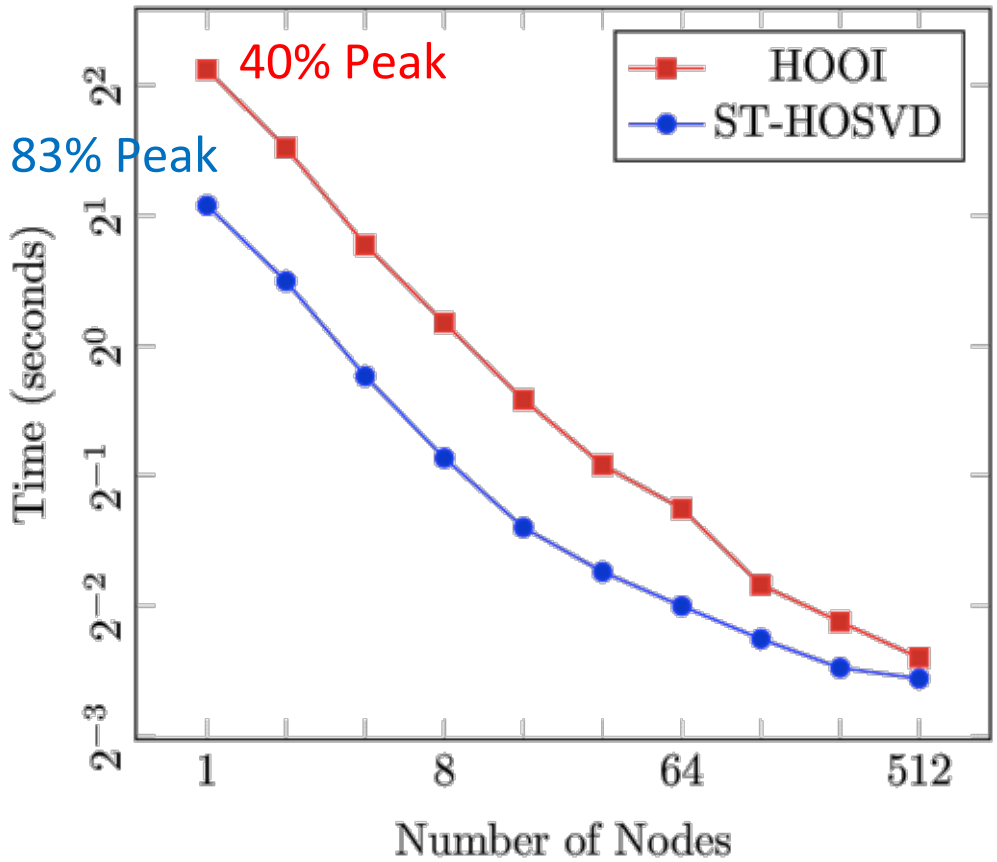


Strong & weak scaling on Edison

Strong Scaling with 24 x 2^k processors

I: 500 x 300 x 240 x 35

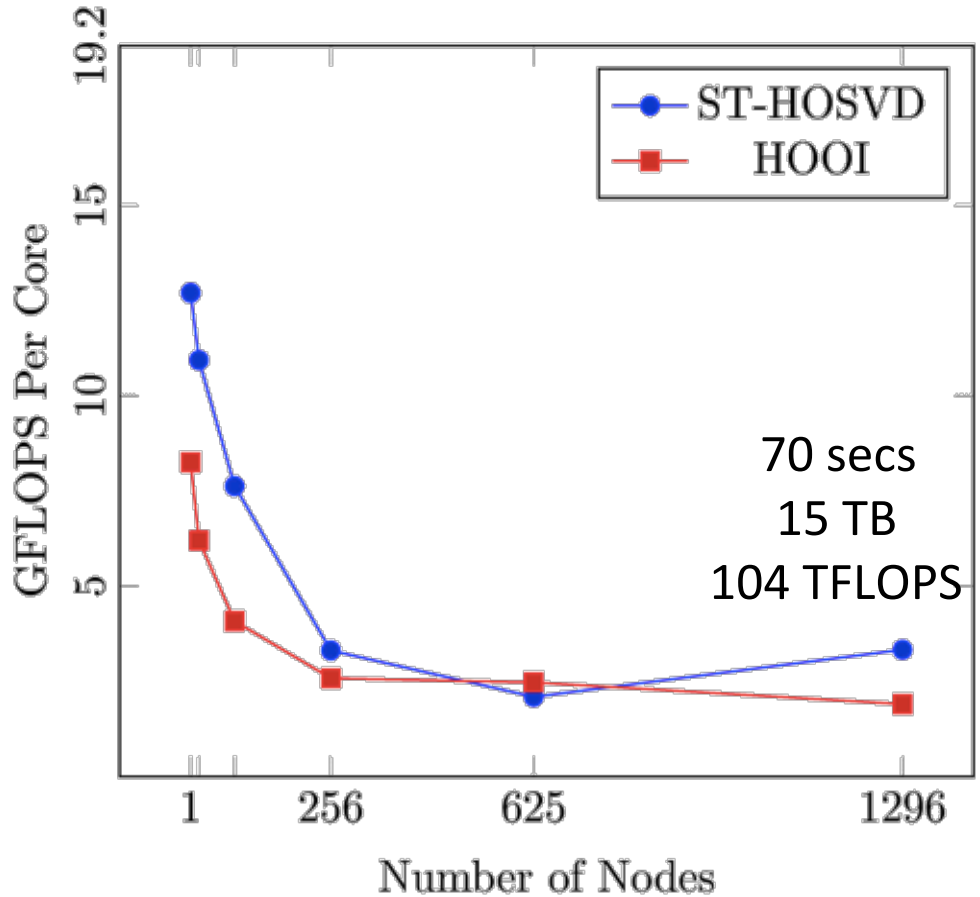
R: 42 x 115 x 81 x 19



Weak Scaling with 24 x k^4 processors

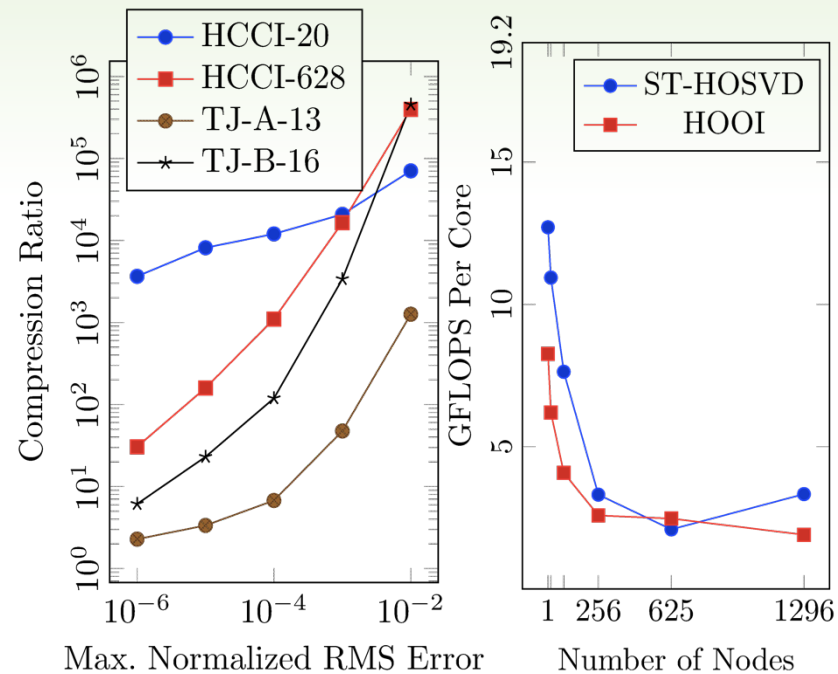
I: 200k x 200k x 200k x 200k

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



Parallel Tucker Compression

- First-ever implementation of distributed-memory parallel Tucker decomposition
- Up to 10^6 compression on real-world data with minimal loss in accuracy
- Scales well – achieving 17% of peak on over 30,000 cores
- Future work
  - Detailed application studies
  - Use QR/SVD instead of Gram/EVD



For more information:
Grey Ballard,
gmballa@sandia.gov

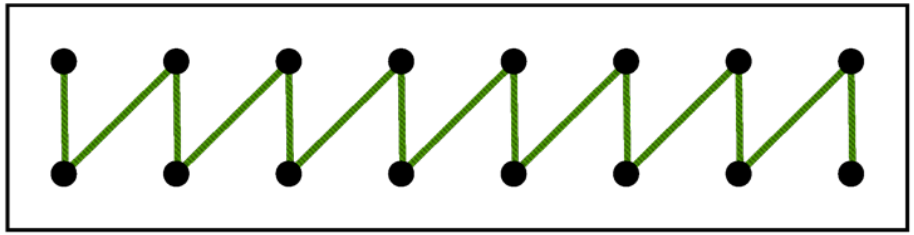
W. Austin, G. Ballard, and T. G. Kolda,
Parallel Tensor Compression for Large-Scale Scientific Data,
http://arxiv.org/abs/1510.06689, October 2015

Backup Slides

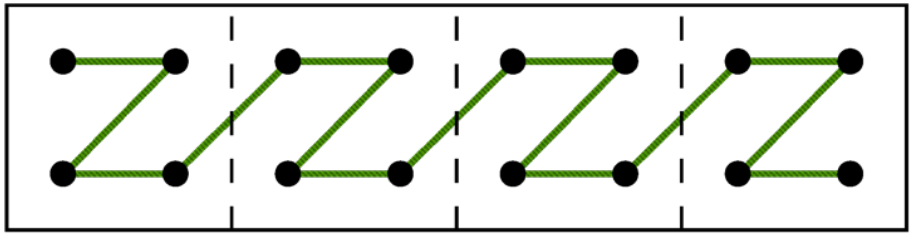
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Local unfolded tensor layout

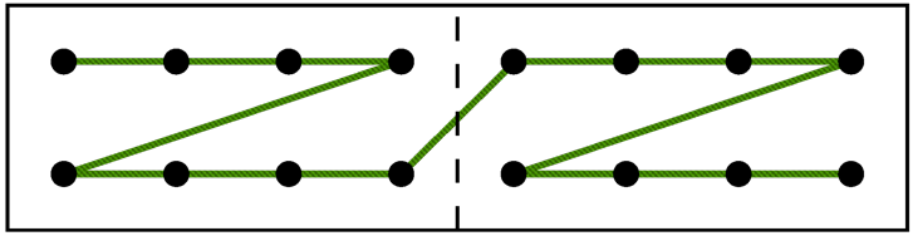
Local Layout: 2 x 2 x 2 x 2



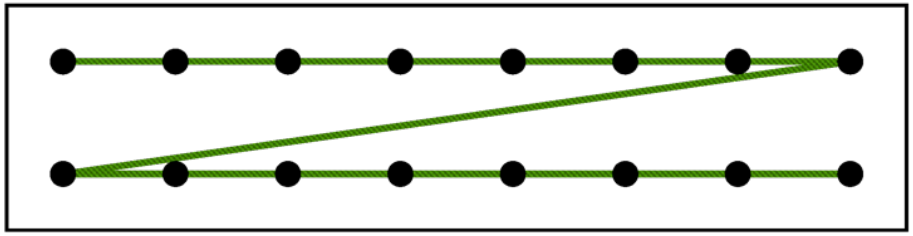
$n = 1$



$n = 2$



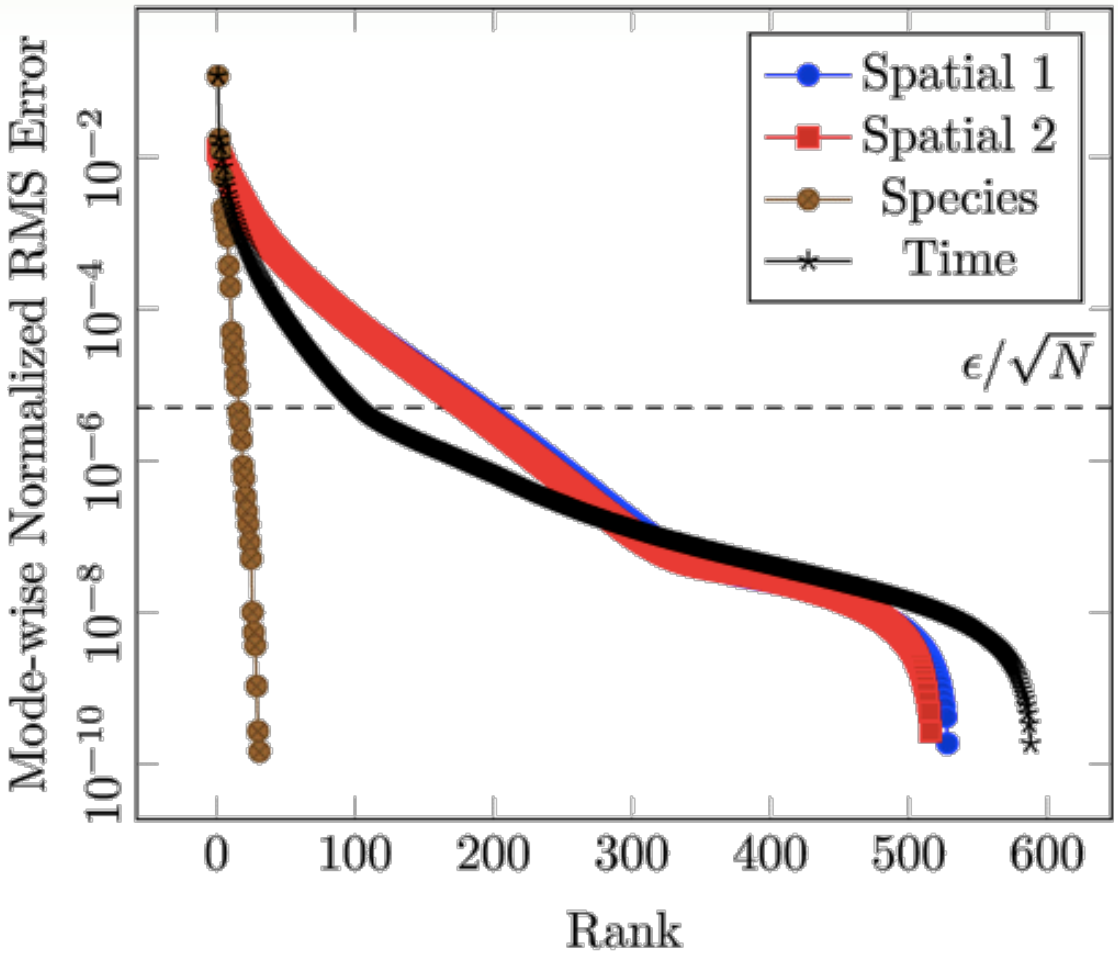
$n = 3$



$n = 4$

Mode-wise contributions to approximation error bound

2D HCCI Data (628 time steps)



$$672 \times 672 \times 33 \times 628 \rightarrow 192 \times 183 \times 16 \times 104$$

Elementwise errors

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

Table 1: Compression and maximum absolute elementwise error of centered data for normalized RMS error of 1e-5.