Performance Prediction for Tensor Contractions

Paolo Bientinesi, Edoardo Di Napoli, Diego Fabregat, Elmar Peise

AICES, RWTH Aachen
pauldj@aices.rwth-aachen.de

June 3rd, 2014
PASCConference 14
Zürich, Switzerland

Tensors – Crash course (1/2)

MATHEMATICS, PHYSICS	$\rightarrow$	multilinear map multidimensional array + metric
COMPUTER SCIENCE	$\rightarrow$	multidimensional array

Tensors – Crash course (1/2)

MATHEMATICS, PHYSICS $\rightarrow$ multilinear map
multidimensional array + metric

COMPUTER SCIENCE $\rightarrow$ multidimensional array

t dimensional tensors $\rightarrow$ $S_{\alpha\beta\ldots\gamma\delta}$, $S_{\underbrace{\alpha\ \ldots\ \delta}_{t \text{ indices}}}^{\beta\quad\gamma}$, $S_{\alpha\beta}{}^{\ldots\gamma\delta}$, $\ldots$

Tensors – Crash course (1/2)

MATHEMATICS, PHYSICS $\rightarrow$ multilinear map
multidimensional array + metric

COMPUTER SCIENCE $\rightarrow$ multidimensional array

t dimensional tensors $\rightarrow$ $S_{\alpha\beta\ldots\gamma\delta}$, $S_{\underbrace{\alpha \quad \ldots \quad \delta}_{t \text{ indices}}}^{\beta \quad \gamma}$, $S_{\alpha\beta}{}^{\ldots\gamma\delta}$, $\ldots$

Operations $\rightarrow$
- low dimensional approximations
- contractions

Tensors – Crash course (1/2)

MATHEMATICS, PHYSICS $\rightarrow$ multilinear map
multidimensional array + metric

COMPUTER SCIENCE $\rightarrow$ multidimensional array

t dimensional tensors $\rightarrow$ $S_{\alpha\beta\ldots\gamma\delta}$, $S_{\underbrace{\alpha\ \ldots\ \delta}_{t \text{ indices}}}^{\beta\quad\gamma}$, $S_{\alpha\beta}{}^{\ldots\gamma\delta}$, $\ldots$

Operations $\rightarrow$
- low dimensional approximations
- contractions

Examples: $S_{\alpha ij}T_{ij}$, $S_{\alpha i\beta}M_{ik}T_{k\gamma}$, $S_{\alpha ij}M_{ik}T_{kh}M_{hj}$, $\ldots$

Tensors – Crash course (1/2)

MATHEMATICS, PHYSICS $\rightarrow$ multilinear map
multidimensional array + metric

COMPUTER SCIENCE $\rightarrow$ multidimensional array

t dimensional tensors $\rightarrow$ $S_{\alpha\beta\ldots\gamma\delta}$, $S_{\underbrace{\alpha\ \ldots\ \delta}_{t \text{ indices}}}^{\ \beta\ \ \gamma}$, $S_{\alpha\beta}^{\ \ \ldots\gamma\delta}$, $\ldots$

Operations $\rightarrow$
- low dimensional approximations
- contractions

Examples: $S_{\alpha ij}T_{ij}$, $S_{\alpha i\beta}M_{ik}T_{k\gamma}$, $S_{\alpha ij}M_{ik}T_{kh}M_{hj}$, $\ldots$

Crash course (2/2)

Contraction $\rightarrow$ $S_{\alpha ij}\, T_{j\delta i}$

α, δ **free** indices
i, j **contracted** indices

Crash course (2/2)

Contraction $\rightarrow$ $S_{\alpha ij}\ T_{j\delta i}$

α, δ **free** indices
i, j **contracted** indices

$$V_{\alpha\delta} = S_{\alpha ij} T_{j\delta i} \quad \rightarrow \quad \forall\alpha\ \forall\delta \quad v_{\alpha\delta} = \sum_i \sum_j s_{\alpha ij}\ t_{j\delta i}$$

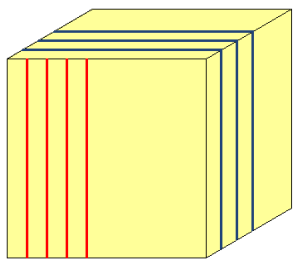
Crash course (2/2)

Contraction $\rightarrow$ $S_{\alpha ij}\ T_{j\delta i}$

α, δ **free** indices
i, j **contracted** indices

$$V_{\alpha\delta} = S_{\alpha ij} T_{j\delta i} \rightarrow \forall\alpha\ \forall\delta \quad v_{\alpha\delta} = \sum_i \sum_j s_{\alpha ij}\ t_{j\delta i}$$

Storage $S_{\alpha\beta\gamma\ldots}$ $\rightarrow$



$\alpha \rightarrow$ stride 1
$\beta \rightarrow$ stride $|\alpha|$
$\gamma \rightarrow$ stride $|\alpha||\beta|$
$\vdots$

A well known contraction (1/2)

$$C_{ij} = A_{ik} B_{kj}$$

A well known contraction (1/2)

$$C_{ij} = A_{ik} B_{kj}$$

0. **Direct call** `C := GEMM(A,B)`

A well known contraction (1/2)

$$C_{ij} = A_{ik} B_{kj}$$

0. **Direct call**

```
C := GEMM(A,B)
```

1. A **is sliced horizontally**

$$C := AB = \left[\begin{array}{c} a^1 \\ \hline \vdots \\ \hline a^m \end{array}\right] B = \left[\begin{array}{c} a^1 B \\ \hline \vdots \\ \hline a^m B \end{array}\right]$$

```
for i=1,...,
  Ci:=GEMV(Ai,B)
```

A well known contraction (1/2)

$$C_{ij} = A_{ik} B_{kj}$$

0. **Direct call**

```
C := GEMM(A,B)
```

1. A **is sliced horizontally**

$$C := AB = \left[\begin{array}{c} a^1 \\ \hline \vdots \\ \hline a^m \end{array}\right] B = \left[\begin{array}{c} a^1 B \\ \hline \vdots \\ \hline a^m B \end{array}\right]$$

```
for i=1,...,
  Ci:=GEMV(Ai,B)
```

2. B **is sliced vertically**

$$C := AB = A[b_1|b_2|\ldots|b_n] = [Ab_1|Ab_2|\ldots|Ab_n]$$

```
for i=1,...,
  Ci:=GEMV(A,Bi)
```

A well known contraction (2/2)

$$C_{ij} = A_{ik} B_{kj}$$

3. **A is sliced vertically and B horizontally**

$$[a_1 | \dots | a_k] \begin{bmatrix} \frac{b^1}{\vdots} \\ \frac{}{b^k} \end{bmatrix} = a_1 b^1 + a_2 b^2 + \cdots + a_k b^k$$

```
for i=1,...,
  C+=GER(Ai,Bi)
```

A well known contraction (2/2)

$$C_{ij} = A_{ik} B_{kj}$$

3. **A is sliced vertically and B horizontally**

$$[a_1|\dots|a_k] \begin{bmatrix} b^1 \\ \hline \vdots \\ \hline b^k \end{bmatrix} = a_1 b^1 + a_2 b^2 + \cdots + a_k b^k$$

```
for i=1,...,
  C+=GER(Ai,Bi)
```

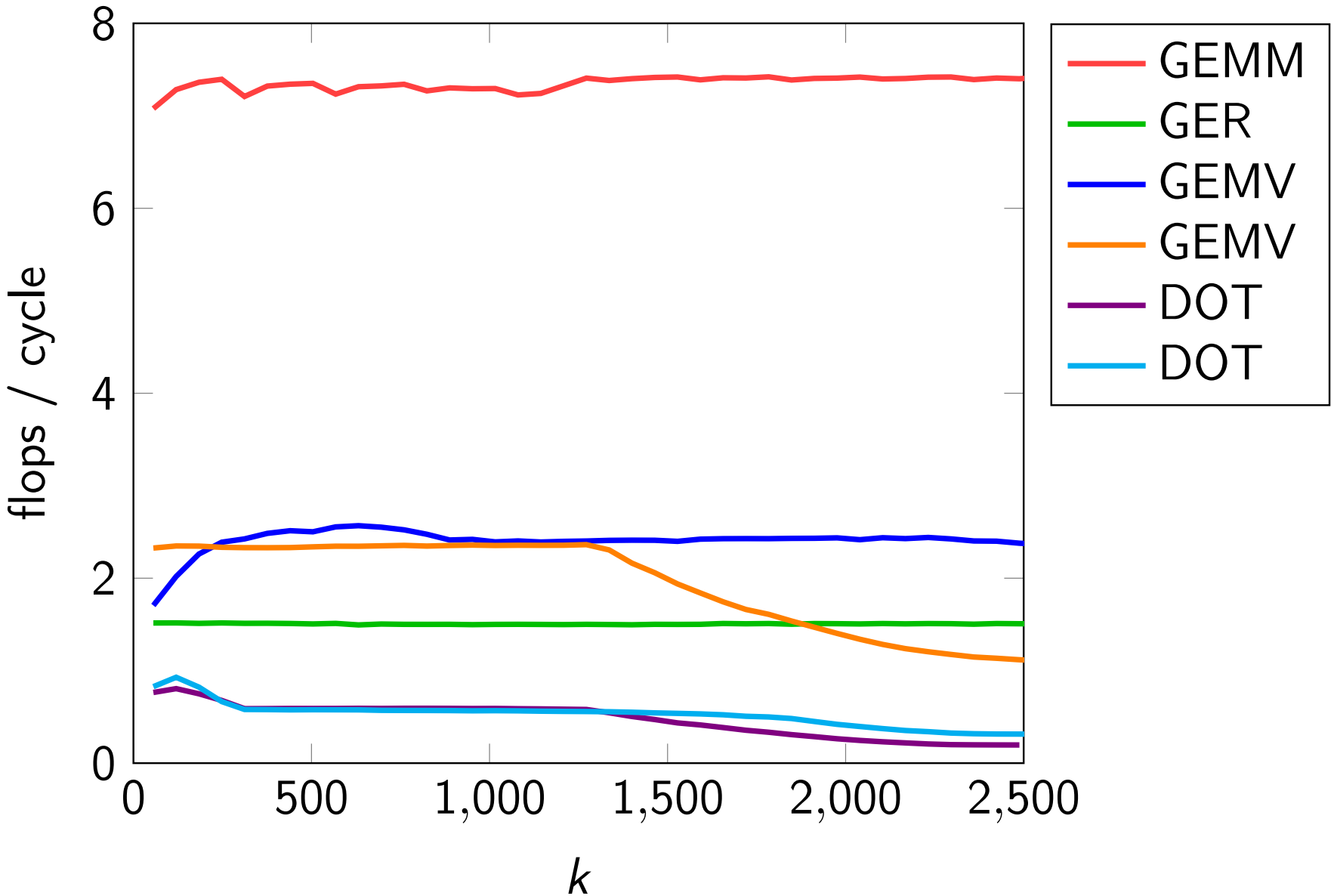
4. **A is sliced horizontally and B vertically**

$$\begin{bmatrix} a^1 \\ \hline \vdots \\ \hline a^m \end{bmatrix} [b_1|\dots|b_n] = \left[\begin{array}{c|c|c} a^1 b_1 & \dots & a^1 b_n \\ \hline \vdots & \ddots & \\ \hline a^m b_1 & & a^m b_n \end{array}\right]$$

```
for i=1,...,
  for j=1,...,
    Cij:=DOT(Ai,Bj)
```

Mathematically equivalent, but . . .

All experiments: OpenBLAS 0.2.8, Intel IvyBridge_EP E5-2680v2



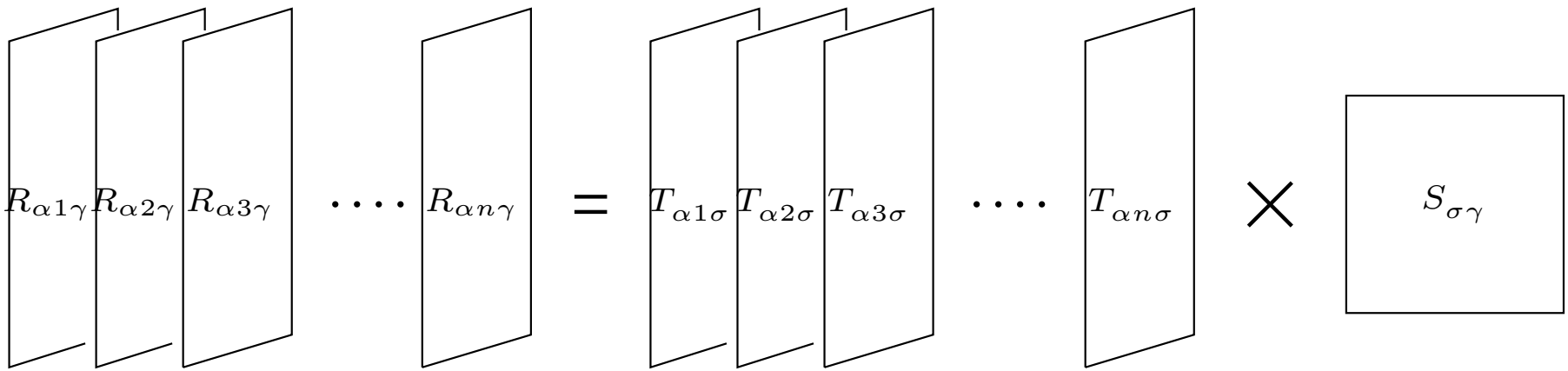
How to use BLAS for contractions?

$$R_{\alpha\beta\gamma} := T_{\alpha\beta\sigma} S_{\sigma\gamma}$$

How to use BLAS for contractions?

$$R_{\alpha\beta\gamma} := T_{\alpha\beta\sigma} S_{\sigma\gamma}$$

1) Slicing along β

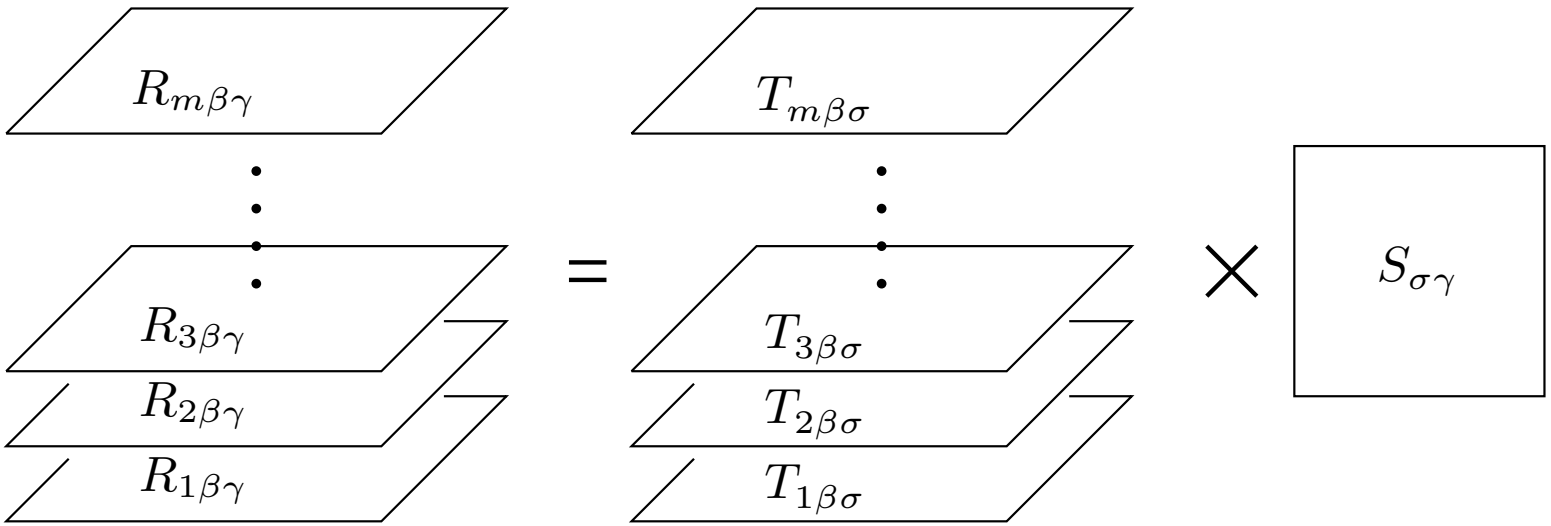


✓ `GEMM`

How to use BLAS for contractions?

$$R_{\alpha\beta\gamma} := T_{\alpha\beta\sigma} S_{\sigma\gamma}$$

2) Slicing along α



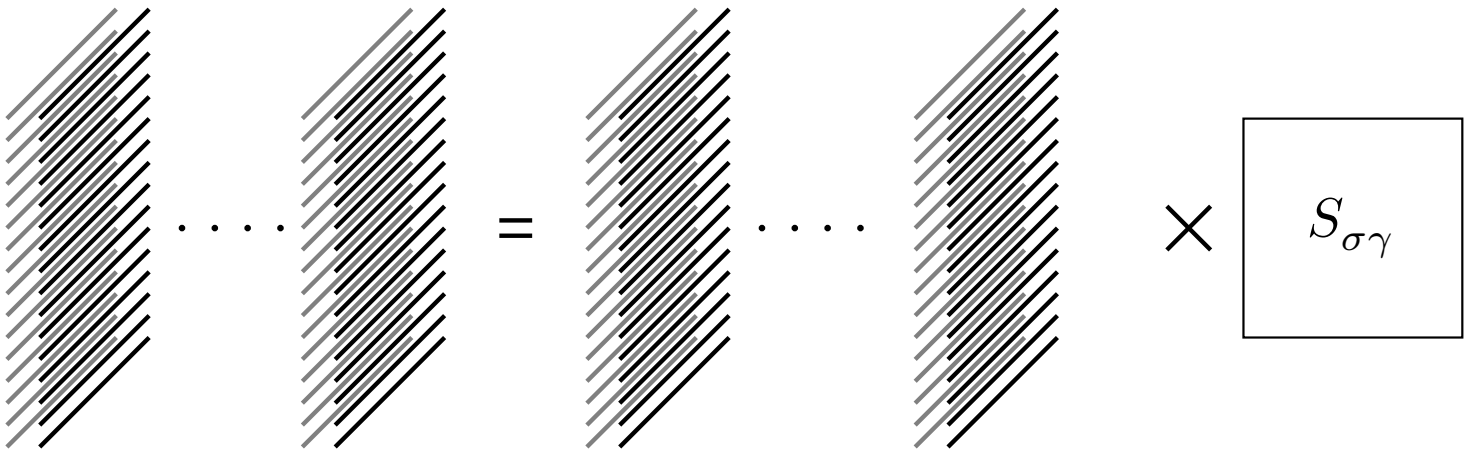
✗ `GEMM`

✓ Transposition + `GEMM`

How to use BLAS for contractions?

$$R_{\alpha\beta\gamma} := T_{\alpha\beta\sigma} S_{\sigma\gamma}$$

3) Slicing along α and β



✓ `GEMV`

Taxonomy

$$V_{h_1 h_2 \ldots} := S_{i_1 i_2 \ldots} T_{j_1 j_2 \ldots}$$

Definition: $\Delta(X) =$ # of free indices of X

Taxonomy

$$V_{h_1 h_2 \ldots} := S_{i_1 i_2 \ldots} T_{j_1 j_2 \ldots}$$

Definition: $\Delta(X) =$ # of free indices of X

Class 1: $\Delta(S) = 0 \ \wedge \ \Delta(T) = 0$

✗ BLAS3 ✗ BLAS2 ✓ BLAS1

Taxonomy

$$V_{h_1 h_2 \ldots} := S_{i_1 i_2 \ldots} T_{j_1 j_2 \ldots}$$

Definition: $\Delta(X) =$ # of free indices of X

Class 1: $\Delta(S) = 0 \ \wedge \ \Delta(T) = 0$

✗ BLAS3 ✗ BLAS2 ✓ BLAS1

Class 2: $\Delta(S) \geq 1 \ \wedge \ \Delta(T) = 0$ or $\Delta(S) = 0 \ \wedge \ \Delta(T) \geq 1$

✗ BLAS3 ✓ BLAS2 (+ transp) ✓ BLAS1

Taxonomy

$$V_{h_1 h_2 \ldots} := S_{i_1 i_2 \ldots} T_{j_1 j_2 \ldots}$$

Definition: $\Delta(X) =$ # of free indices of X

Class 1: $\Delta(S) = 0 \ \wedge \ \Delta(T) = 0$

✗ BLAS3 ✗ BLAS2 ✓ BLAS1

Class 2: $\Delta(S) \geq 1 \ \wedge \ \Delta(T) = 0$ or $\Delta(S) = 0 \ \wedge \ \Delta(T) \geq 1$

✗ BLAS3 ✓ BLAS2 (+ transp) ✓ BLAS1

Class 3: $\Delta(S) \geq 1 \ \wedge \ \Delta(T) \geq 1$

✓ BLAS3 (+ transp) ✓ BLAS2 ✓ BLAS1

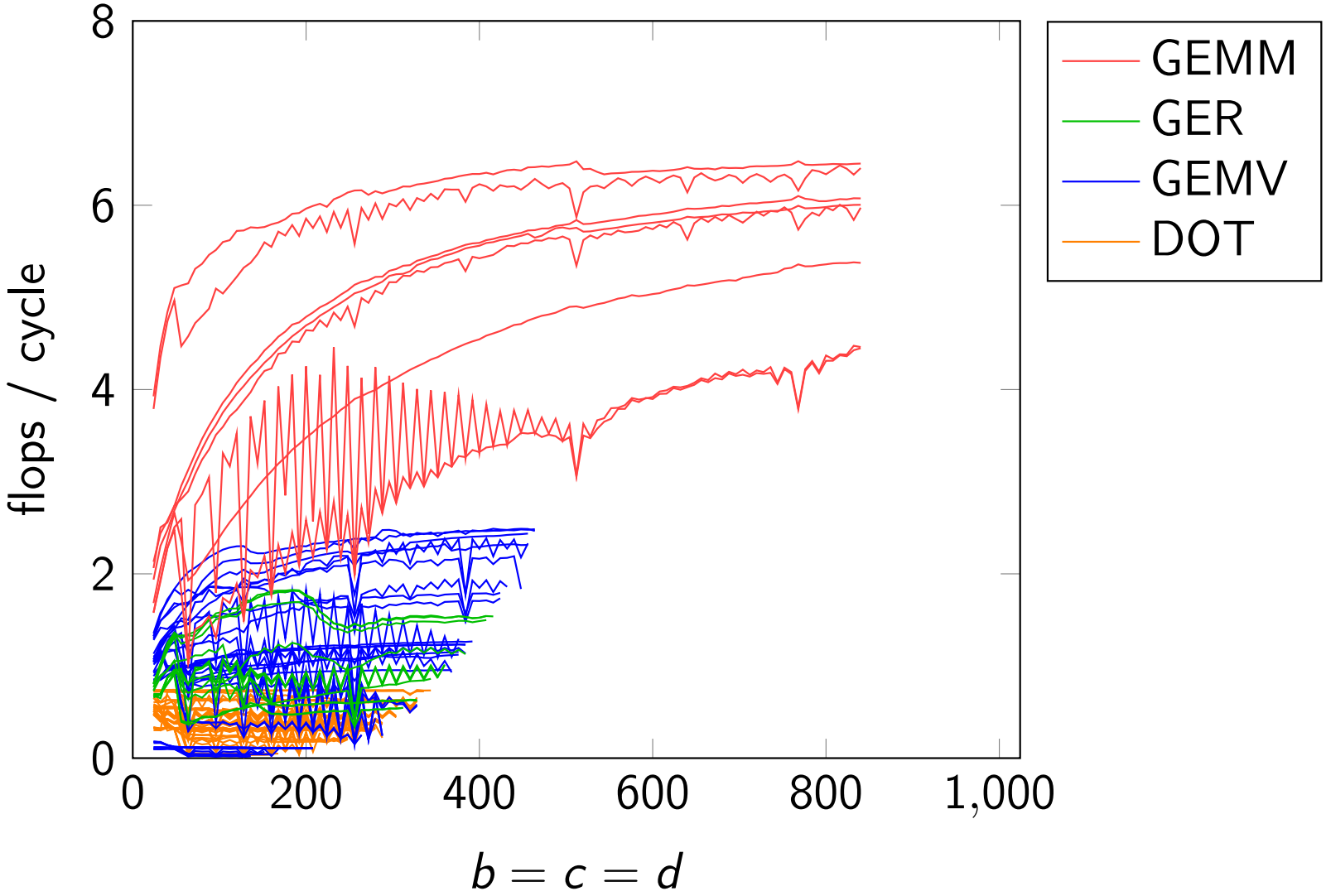
"Towards an Efficient Use of the BLAS Library for Multilinear Tensor Contractions", Applied Mathematics and Computation, 2014.

Nice and easy...

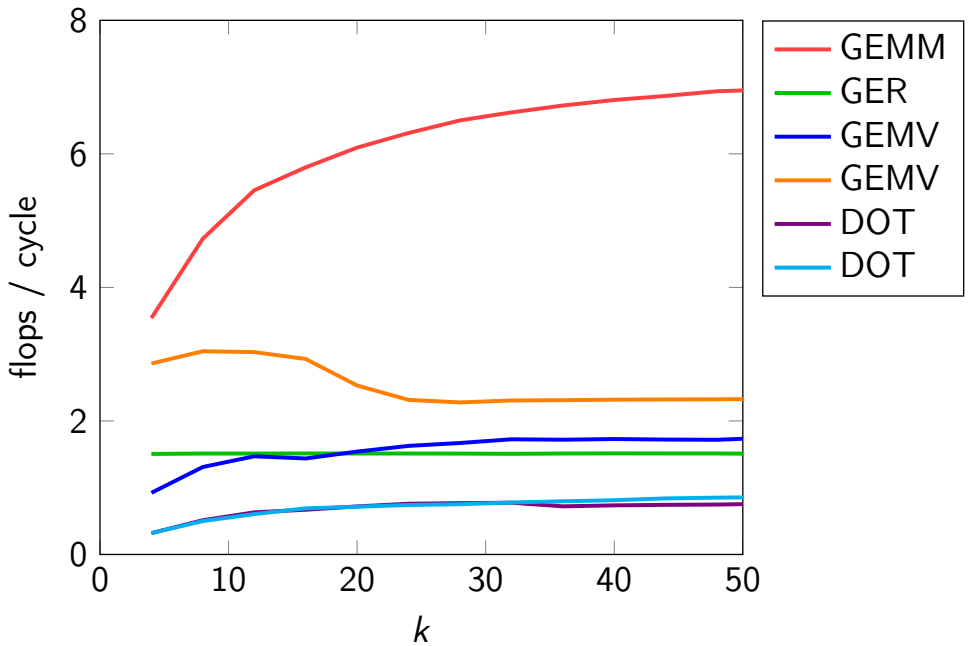
$$V_{bcd} := S_{ijb}\, T_{icjd}$$

Nice and easy... ...104 variants

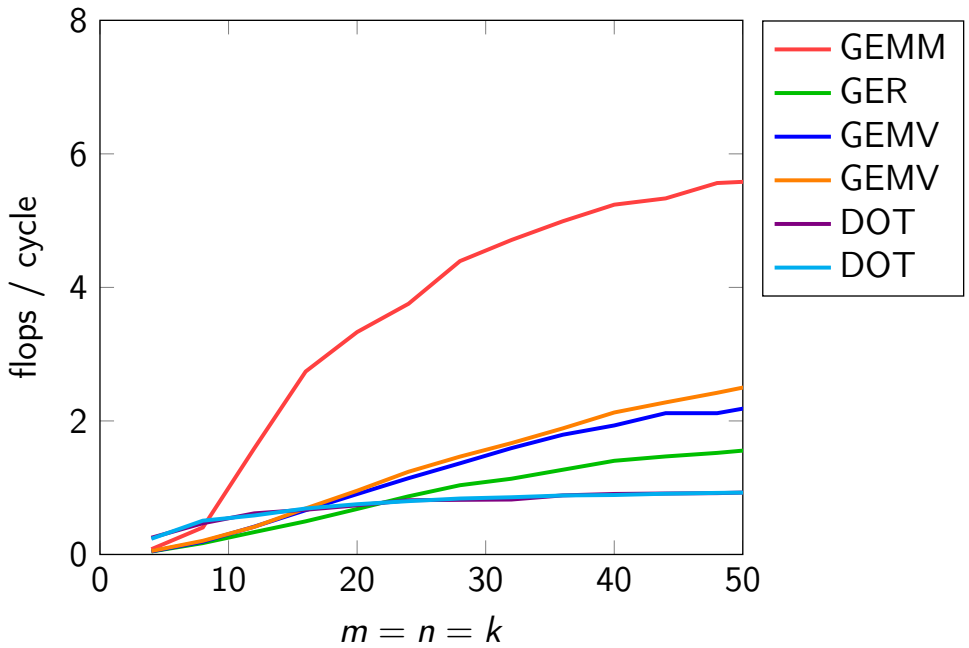
$$V_{bcd} := S_{ijb}\, T_{icjd}$$



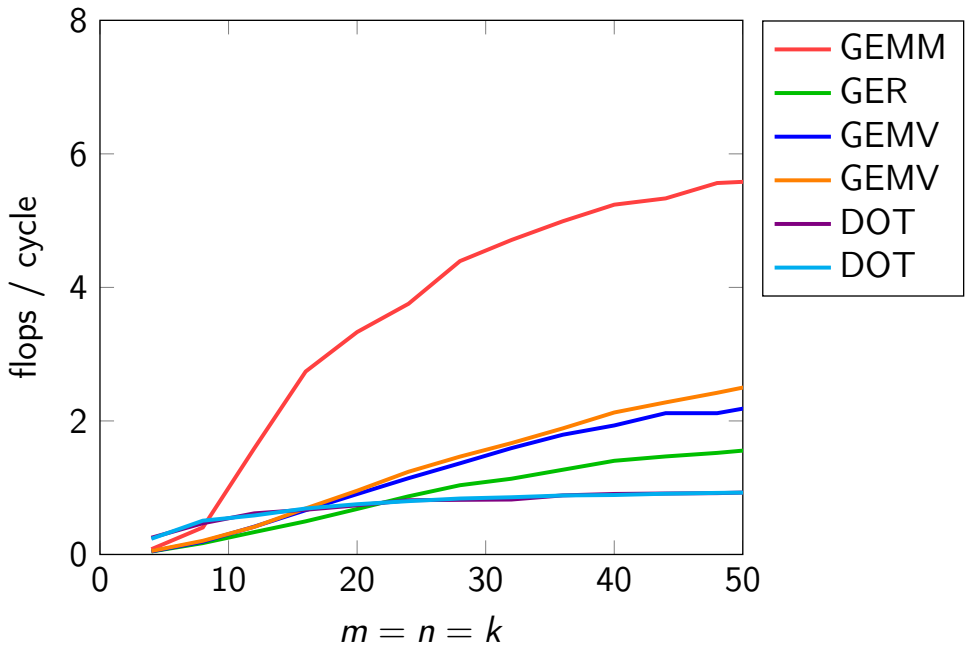
Small dimensions...

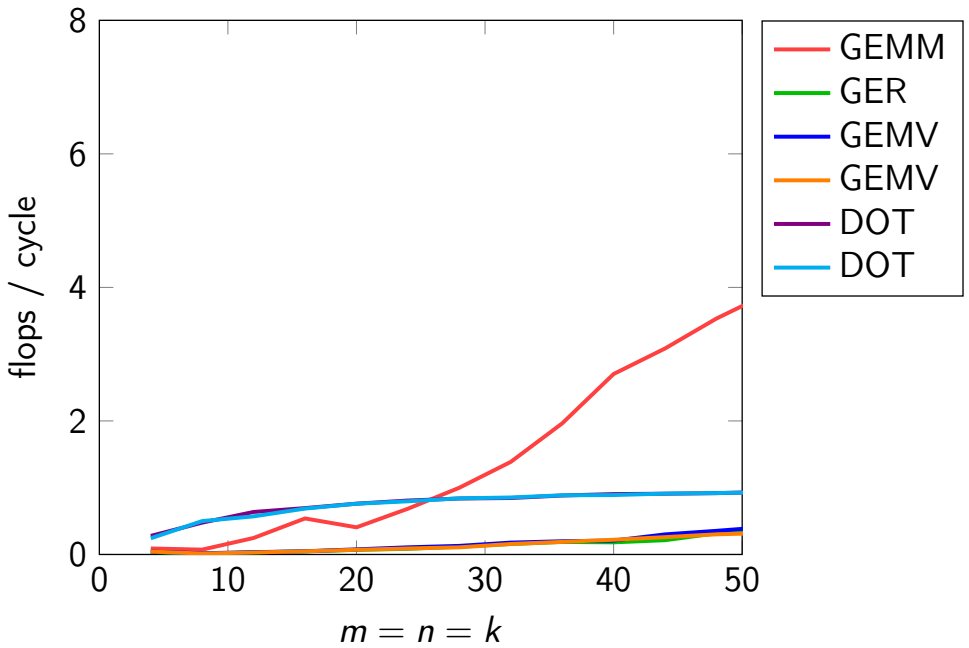


Small dimensions...



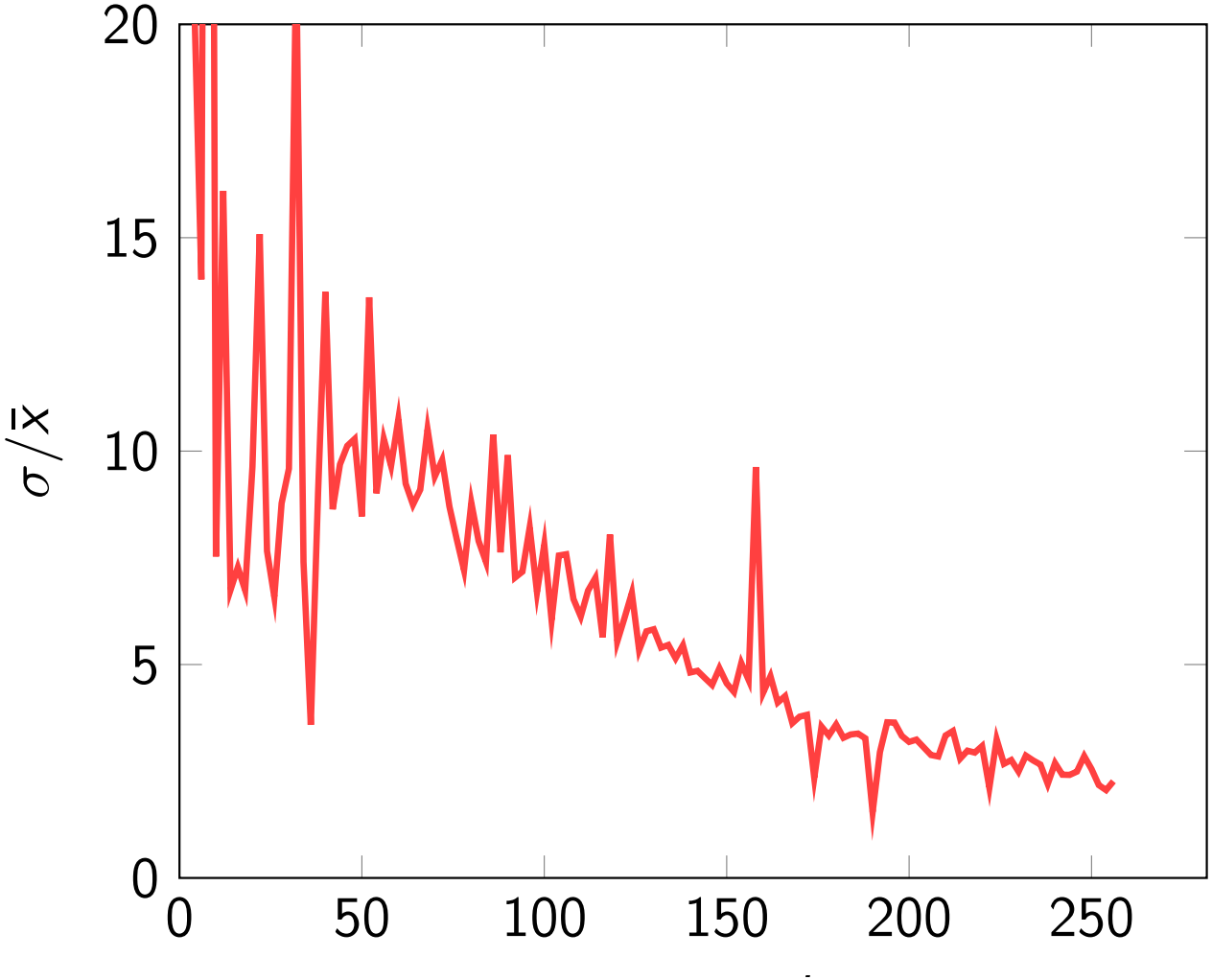
Small dimensions...



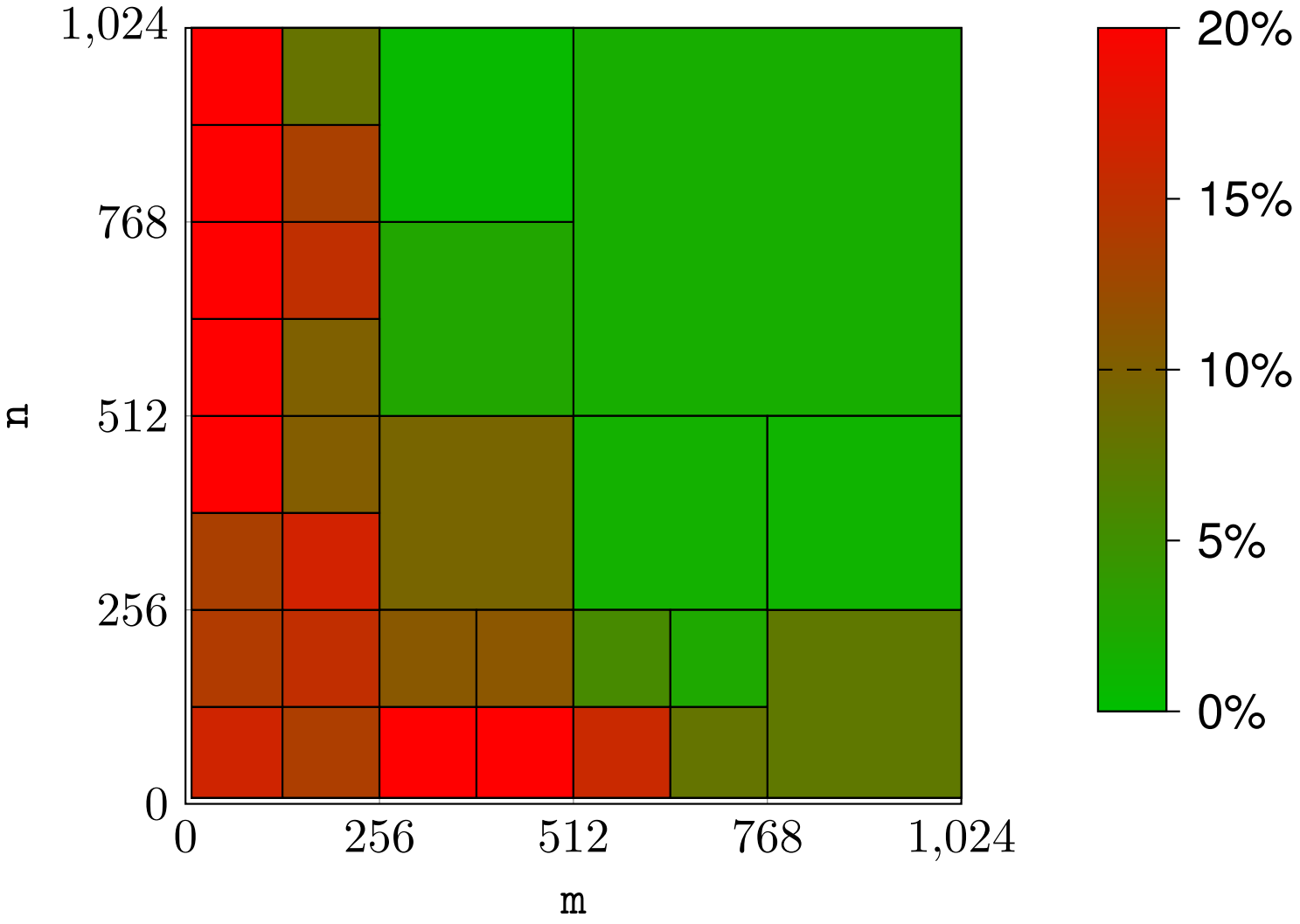


Goal	Automatic **selection** of the best variants
Idea	Performance prediction
Approach	- Kernels execution ✓ - Algorithms execution ✗
Challenges	Fluctuations – uncertainties Cache influence
Solution	- Performance models ✗ - Context-aware timings

Fluctuations $\rightarrow$ performance models

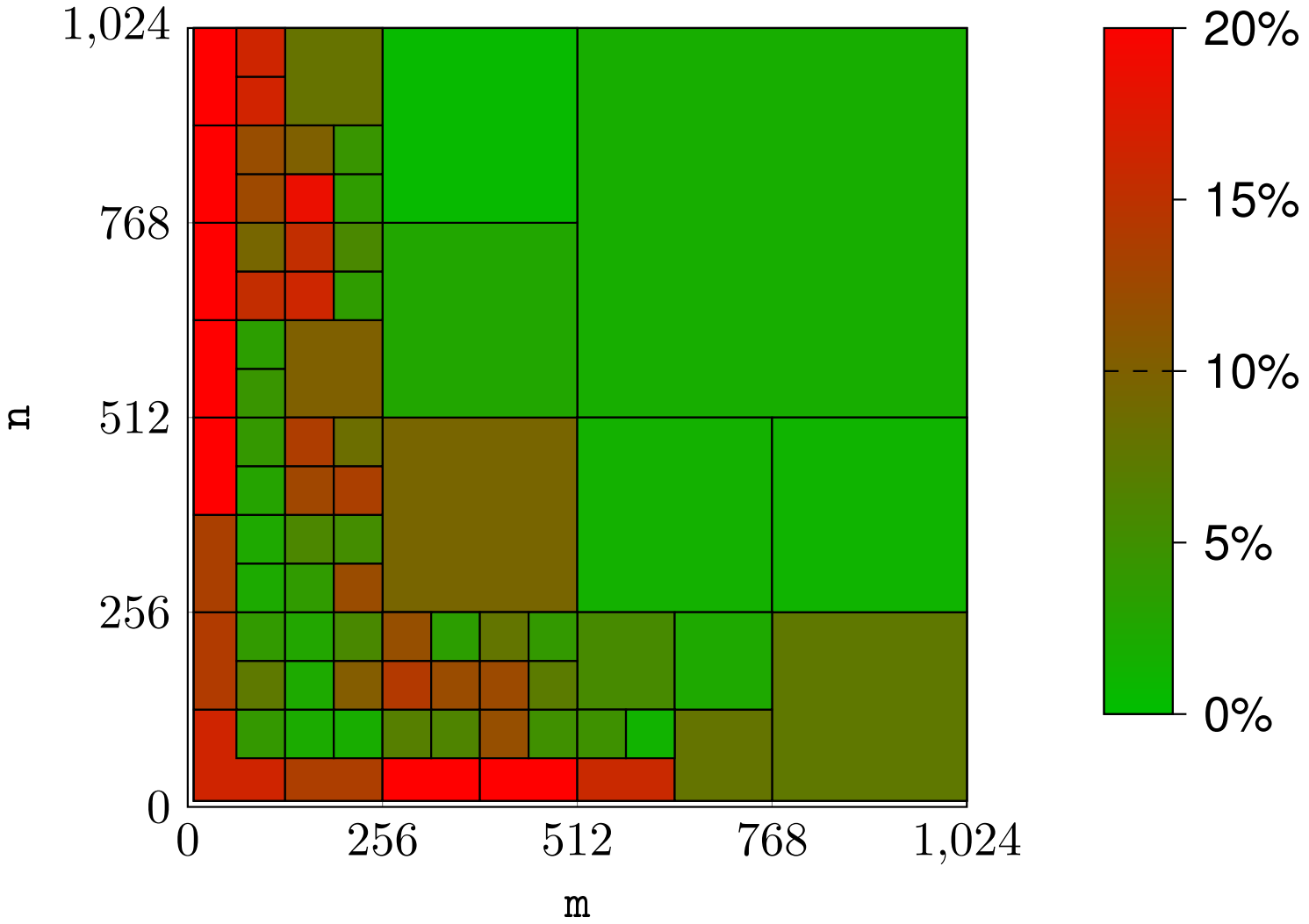


Fluctuations $\rightarrow$ performance models



"Performance Modeling for Dense Linear Algebra", E. Peise, P.B., PMBS12 (SC12).

Fluctuations $\rightarrow$ performance models



"Performance Modeling for Dense Linear Algebra", E. Peise, P.B., PMBS12 (SC12).

~~Models~~ . . . Timings?

Observation:

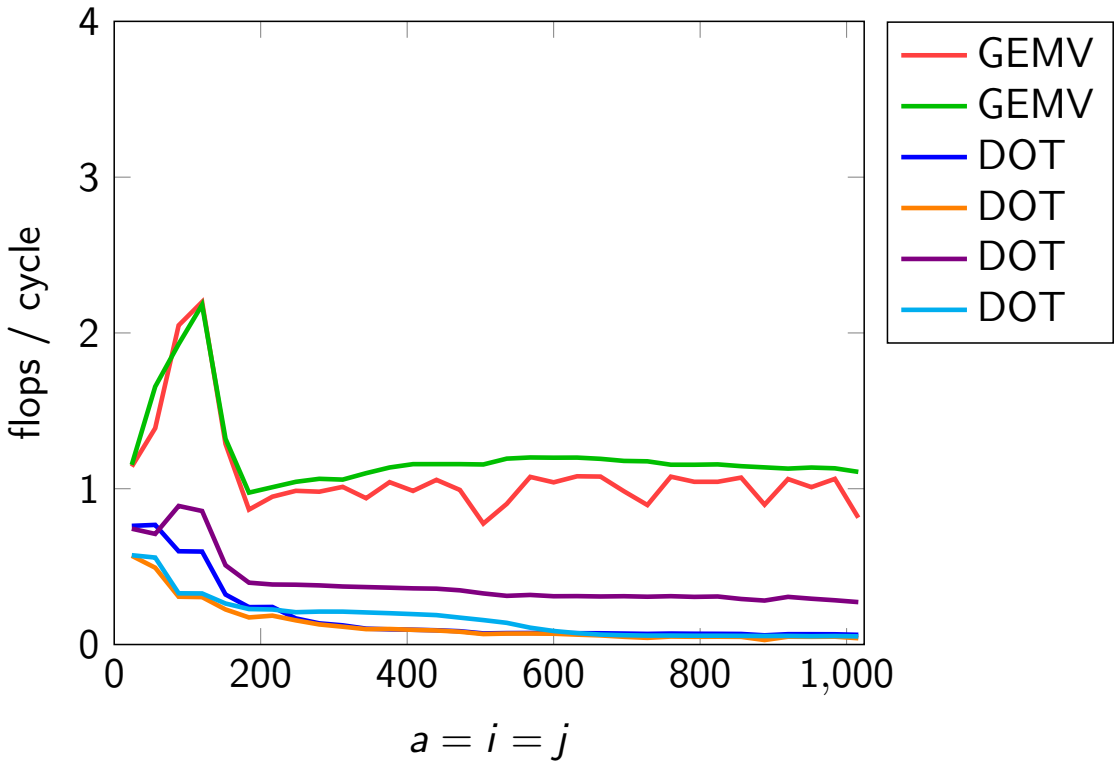
typical linear algebra algorithms	$\rightarrow$	shrinking active region
tensor contractions	$\rightarrow$	identical size slices

~~Models~~ . . . Timings?

Observation:

typical linear algebra algorithms $\rightarrow$ shrinking active region
tensor contractions $\rightarrow$ identical size slices

$$V_a := S_{aij} T_{ij}$$

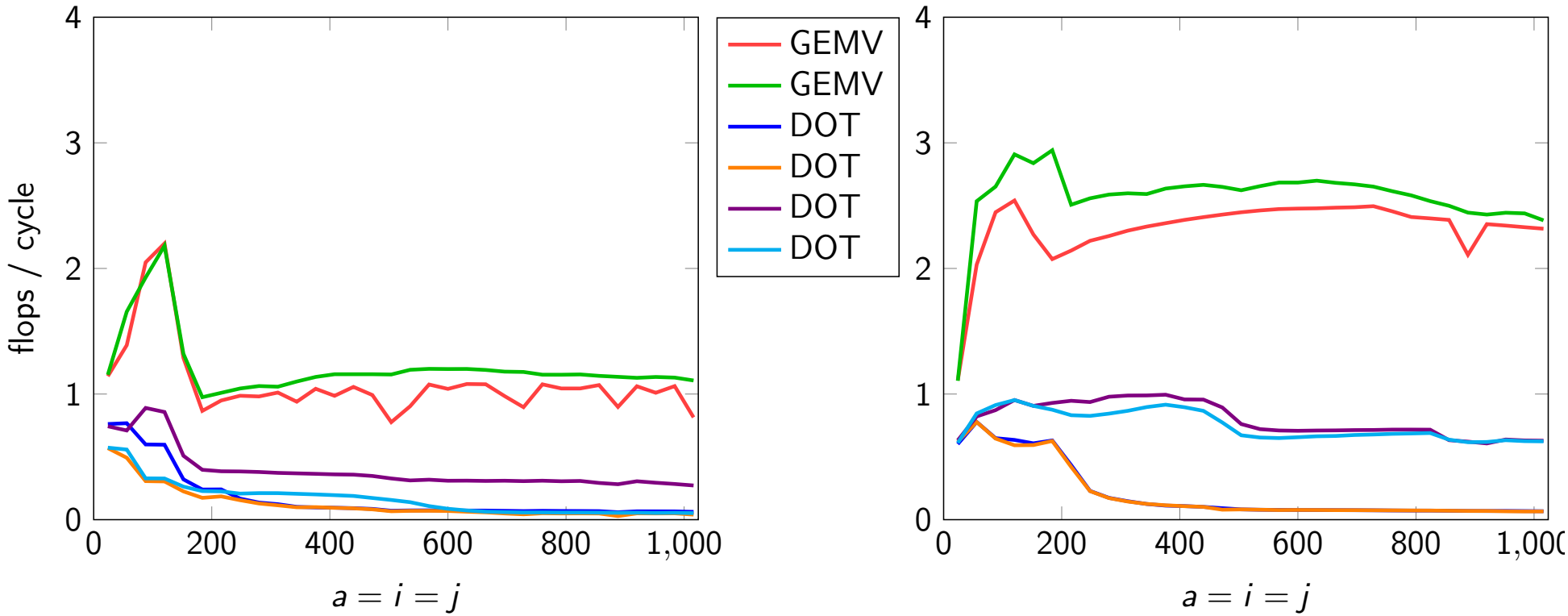


~~Models~~ . . . Timings?

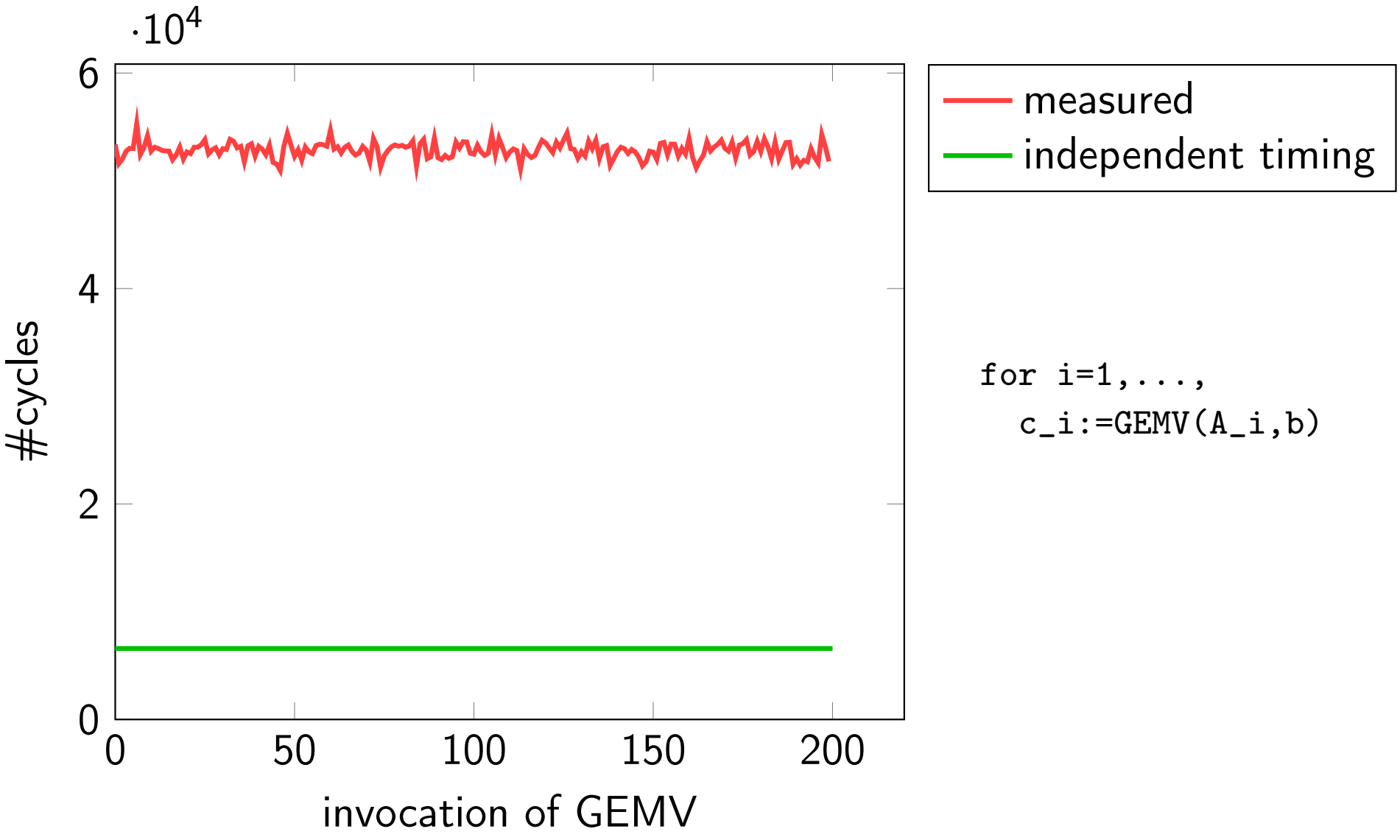
Observation:

typical linear algebra algorithms $\rightarrow$ shrinking active region
tensor contractions $\rightarrow$ identical size slices

$$V_a := S_{aij} T_{ij}$$

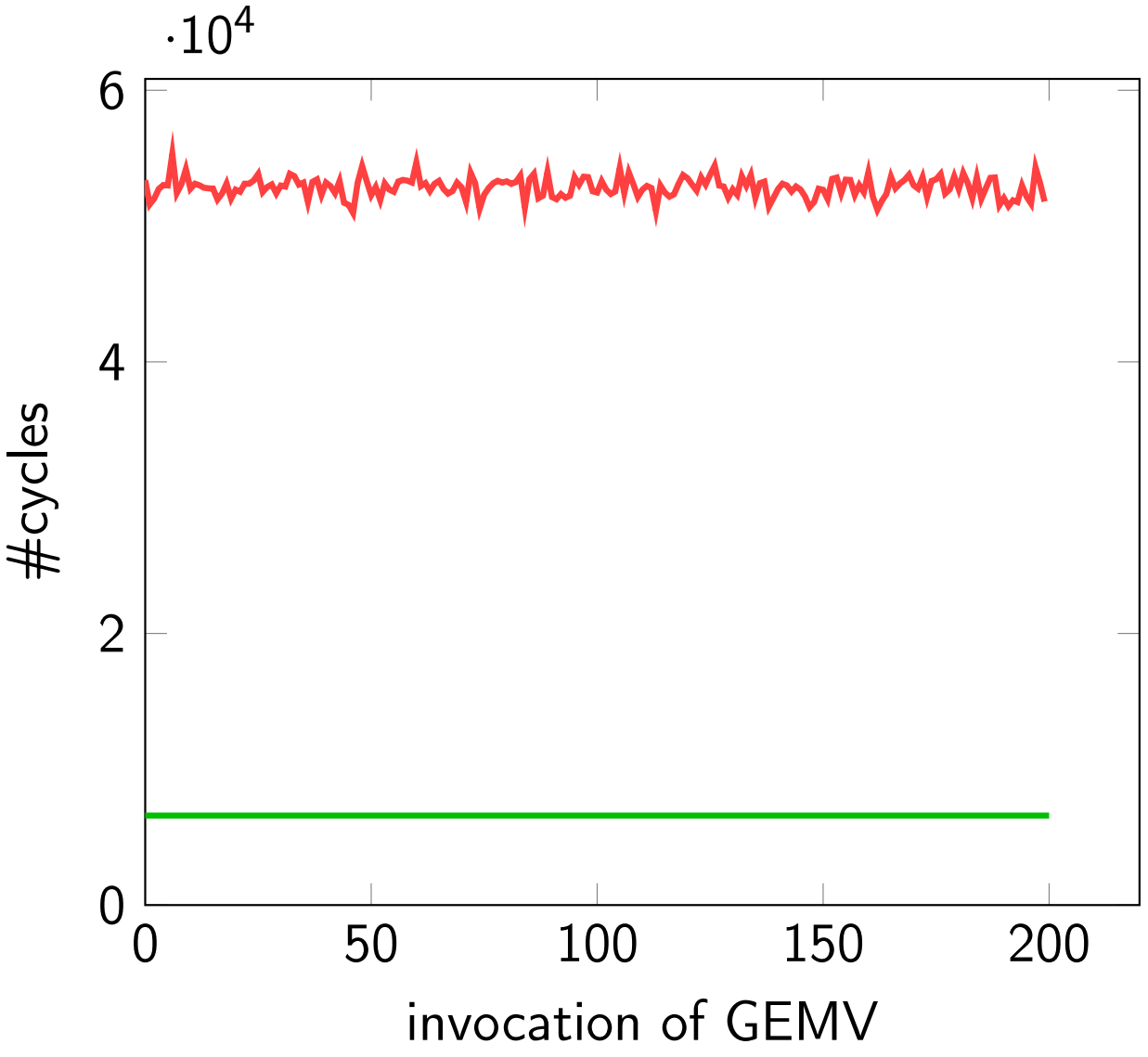


Influence of caching (1/2)




```
for i=1,...,
  c_i:=GEMV(A_i,b)
```

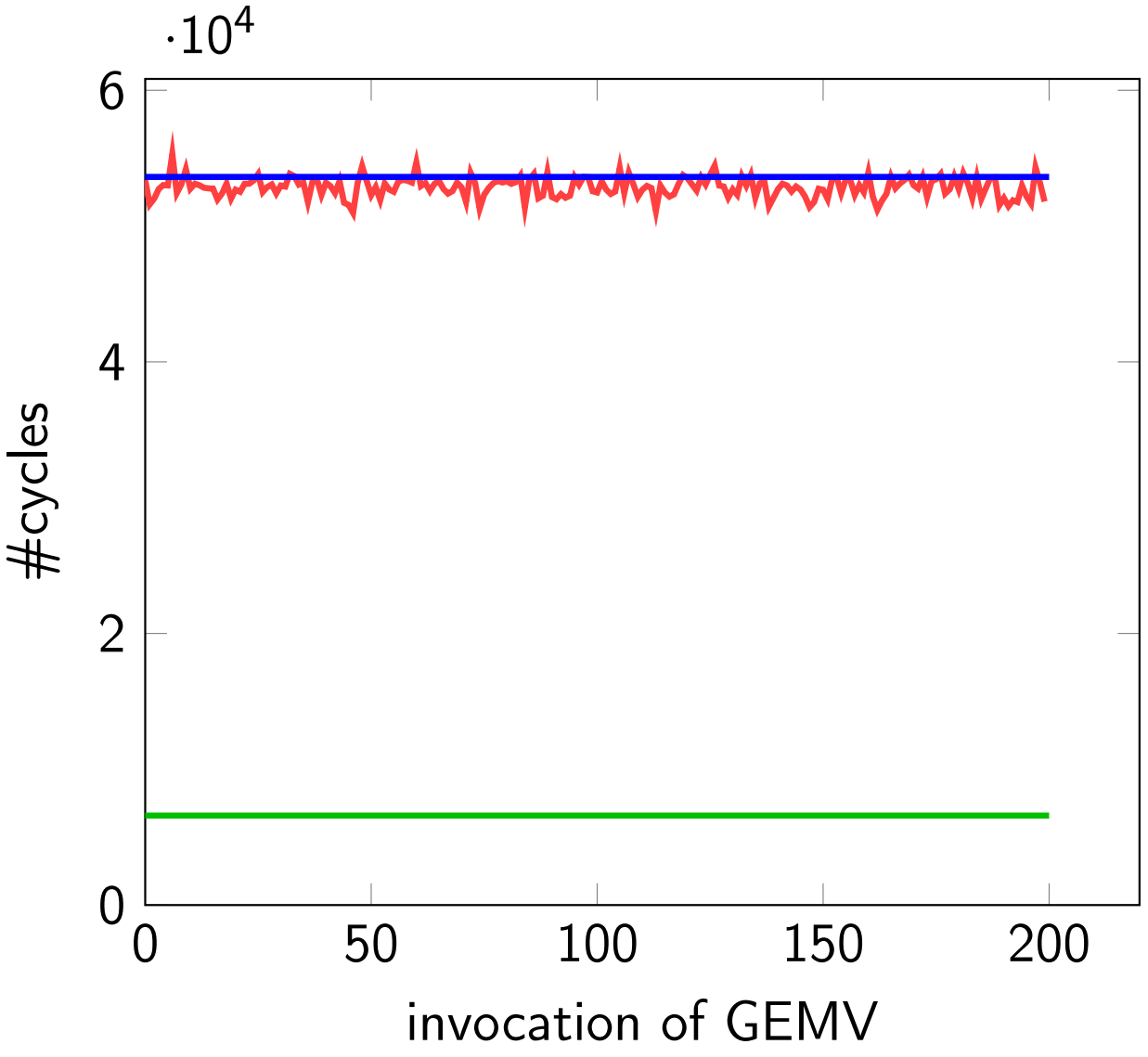
Influence of caching (1/2)




```
for i=1,...,
  c_i:=GEMV(A_i,b)
```

Idea: cache setup

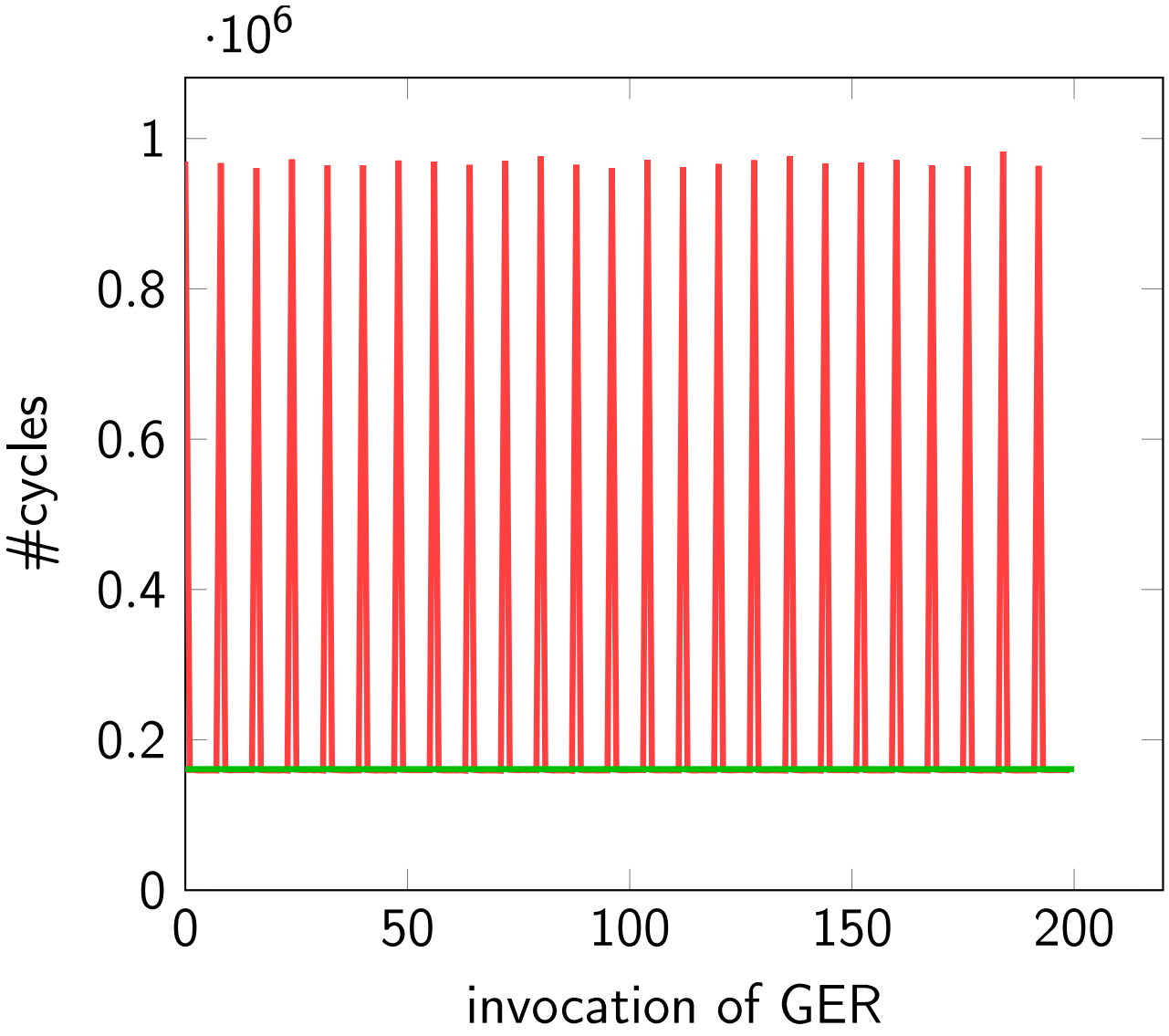
Influence of caching (1/2)




```
for i=1,...,
  c_i:=GEMV(A_i,b)
```

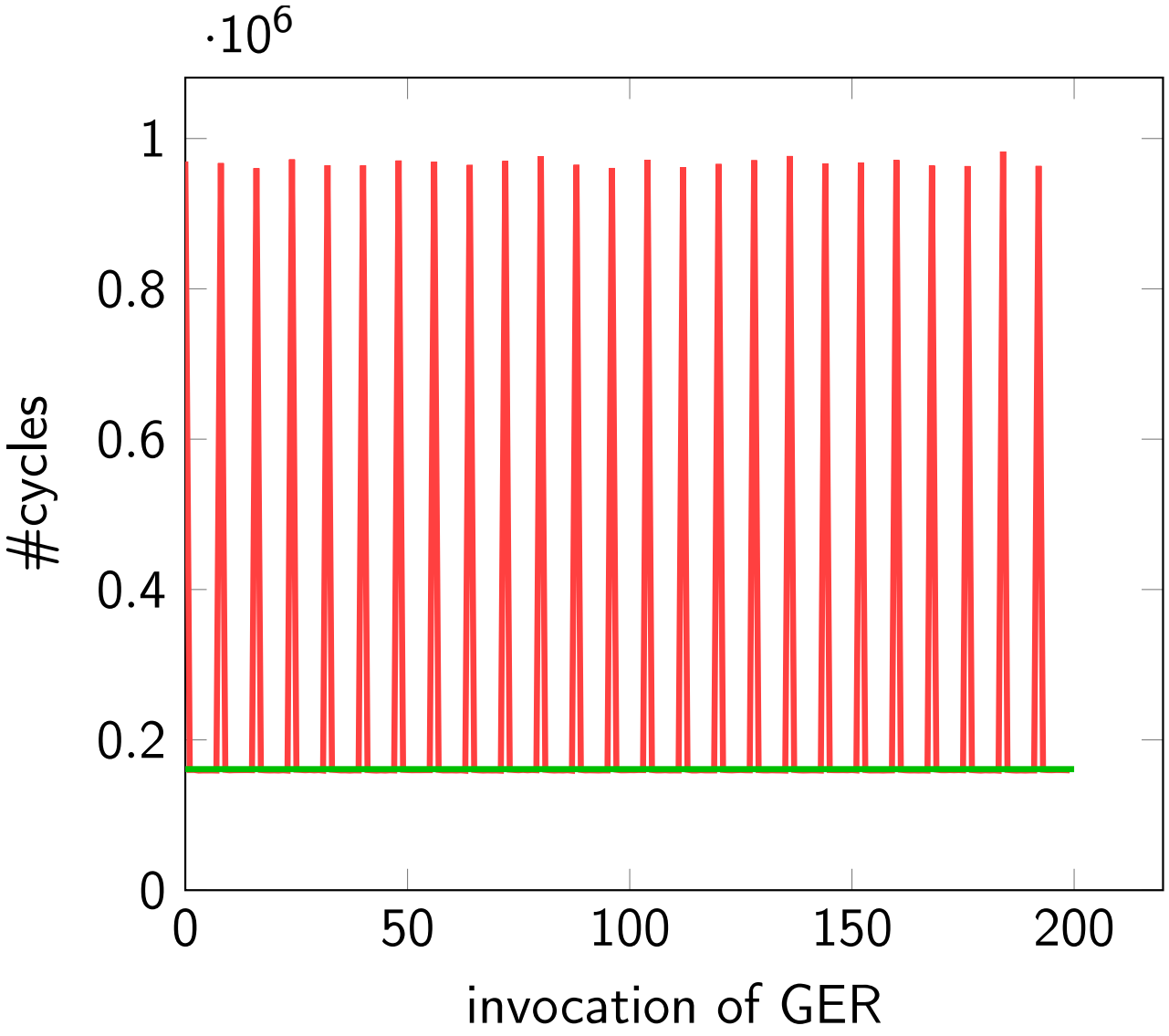
Idea: cache setup

Influence of caching (2/2)




```
for i=1,...,
  for j=1,...,
    A_i:=GER(ai,bj)
```

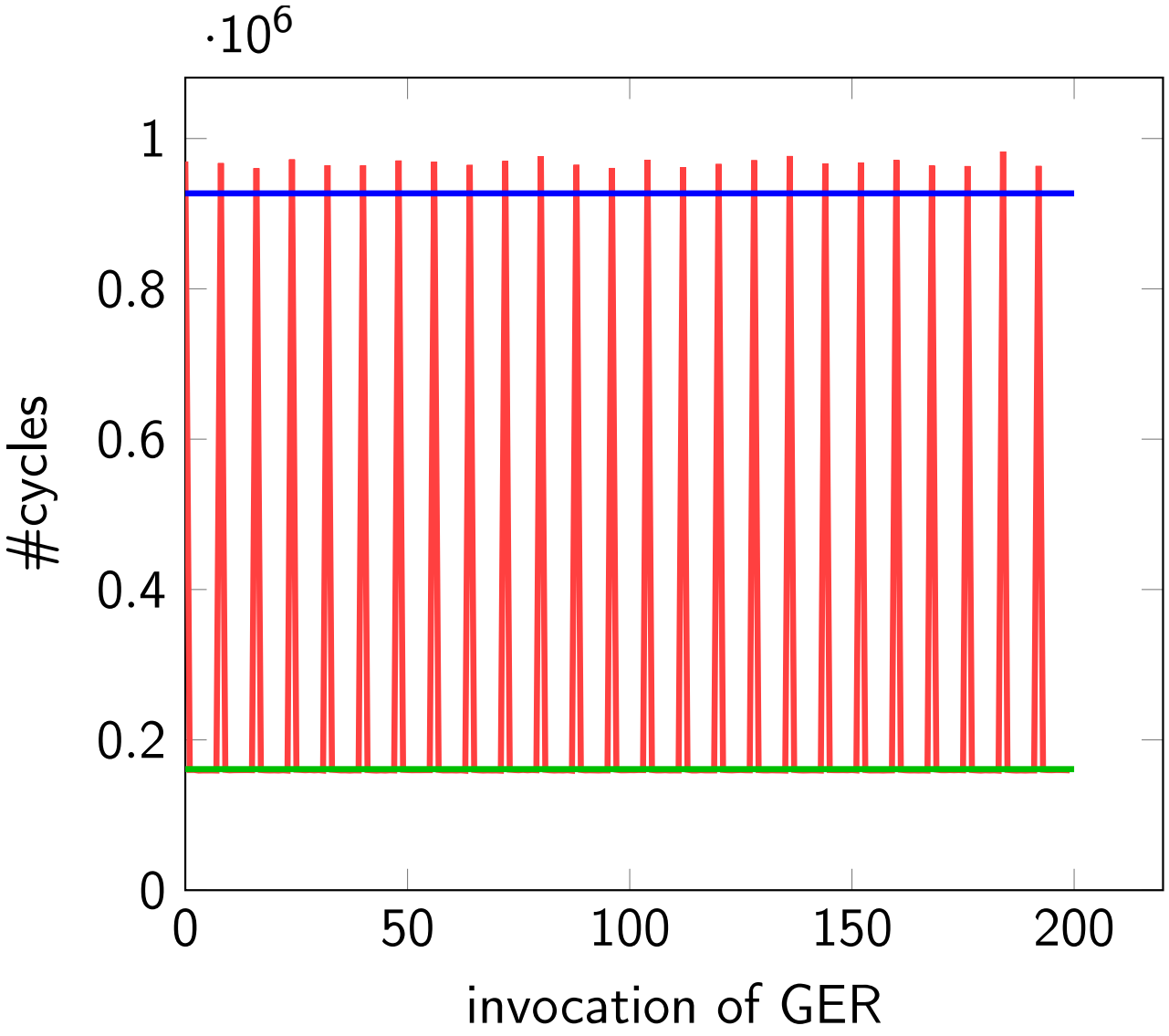
Influence of caching (2/2)




```
for i=1,...,
  for j=1,...,
    A_i:=GER(ai,bj)
```

Idea: first iteration

Influence of caching (2/2)

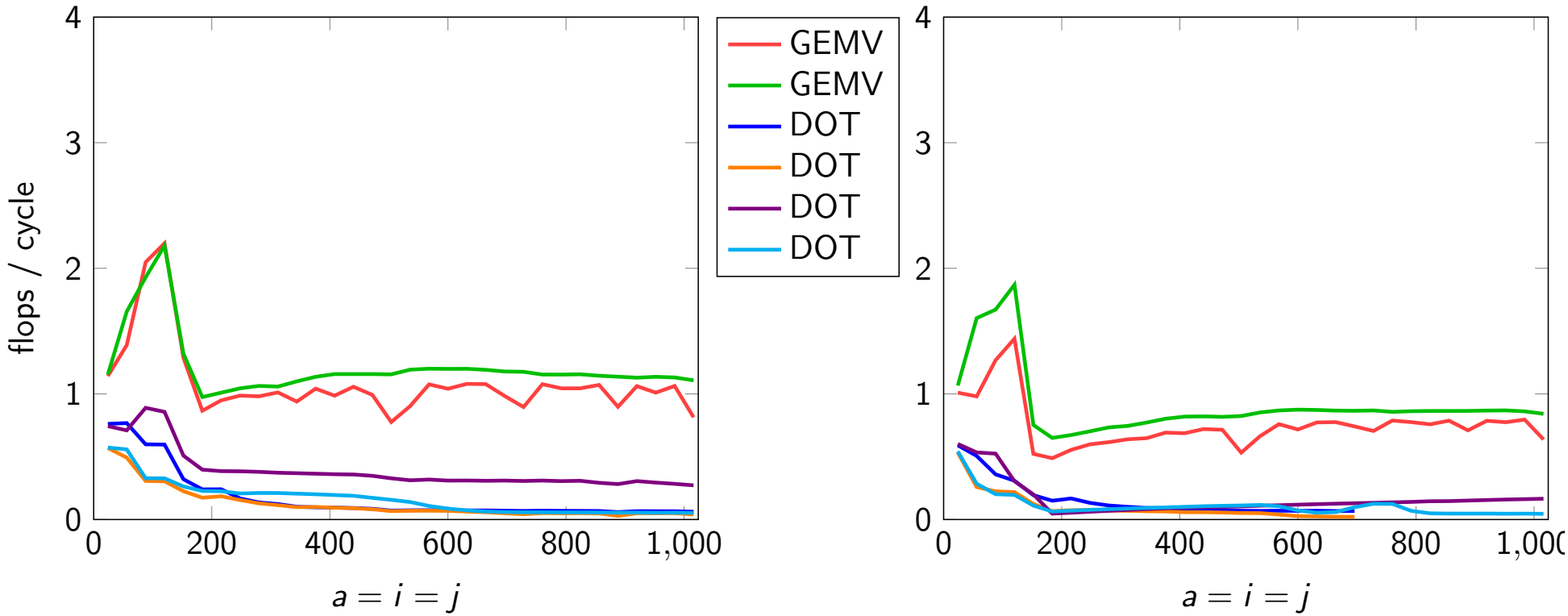



```
for i=1,...,
  for j=1,...,
    A_i:=GER(ai,bj)
```

Idea: first iteration

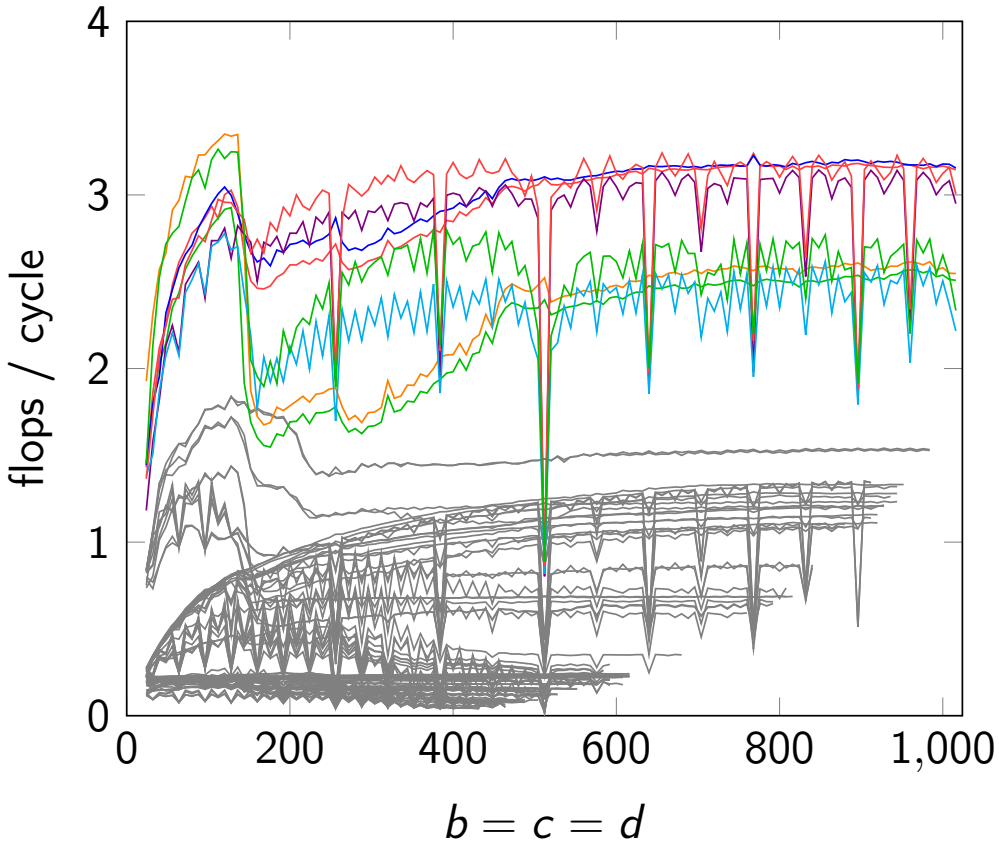
(Nice) results

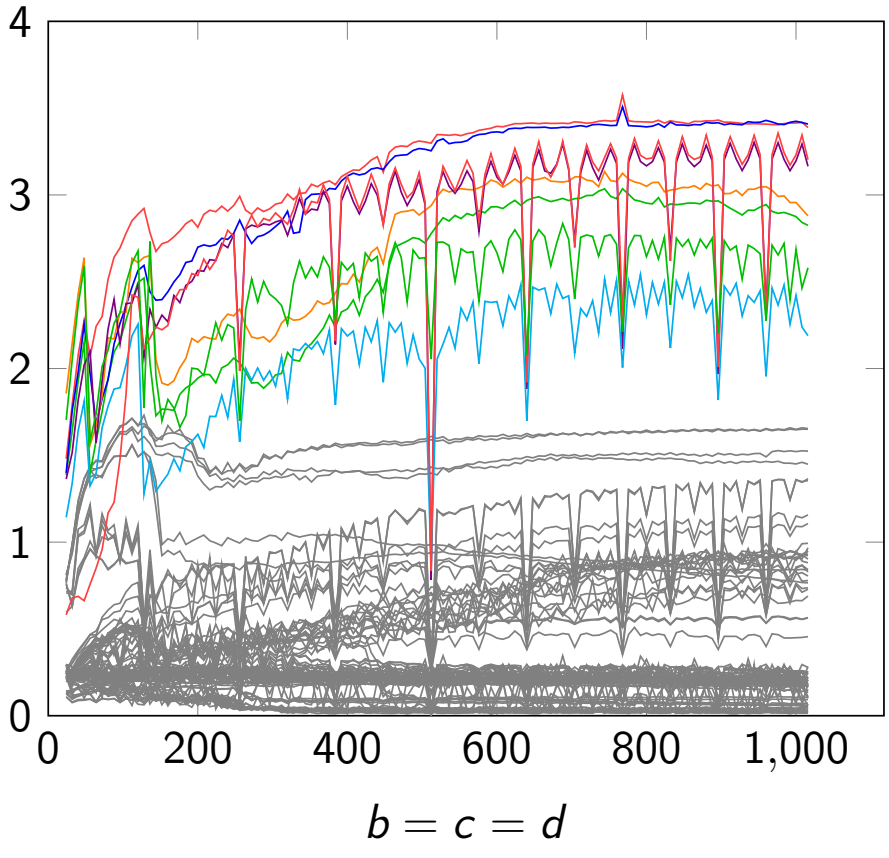
$$V_a := S_{aij} T_{ij}$$



(So so) results

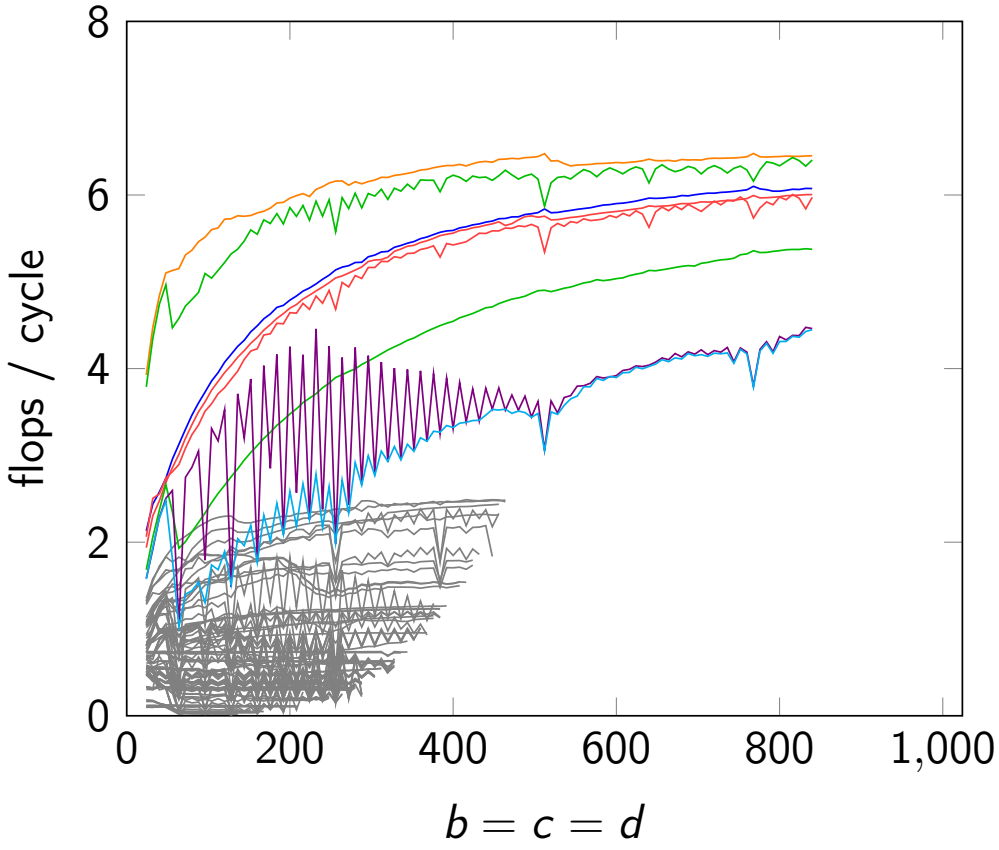
$$V_{bcd} := S_{ijb} T_{icjd}$$

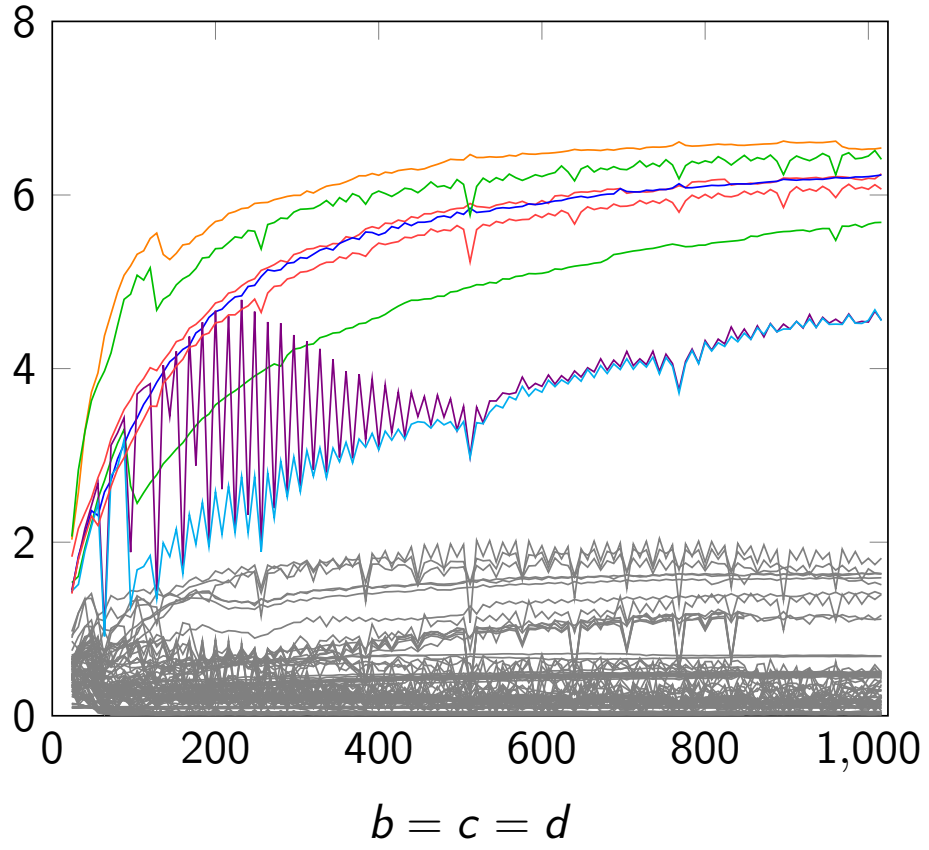




(Awesome) results

$$V_{bcd} := S_{ijb} T_{icjd}$$





Conclusions

Tensor contractions	$\neq$ matrix operations
algorithmic space	LARGE!
need for BLAS4?	maybe
automation?	Yes, please!