



**HAL**  
open science

# Exploring variable accuracy storage through lossy compression techniques in numerical linear algebra: a first application to flexible GMRES

Emmanuel Agullo, Franck Cappello, Sheng Di, Luc Giraud, Xin Liang, Nick Schenkels

## ► To cite this version:

Emmanuel Agullo, Franck Cappello, Sheng Di, Luc Giraud, Xin Liang, et al.. Exploring variable accuracy storage through lossy compression techniques in numerical linear algebra: a first application to flexible GMRES. [Research Report] RR-9342, Inria Bordeaux Sud-Ouest. 2020. hal-02572910v1

**HAL Id: hal-02572910**

**<https://inria.hal.science/hal-02572910v1>**

Submitted on 13 May 2020 (v1), last revised 9 Jul 2020 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# Exploring variable accuracy storage through lossy compression techniques in numerical linear algebra: a first application to flexible GMRES

Emmanuel Agullo, Franck Cappello, Sheng Di, Luc Giraud, Xin  
Liang, Nick Schenkels

**RESEARCH  
REPORT**

**N° 9342**

May 2020

Project-Team HiePACS





## Exploring variable accuracy storage through lossy compression techniques in numerical linear algebra: a first application to flexible GMRES

Emmanuel Agullo<sup>\*</sup>, Franck Cappello<sup>†</sup>, Sheng Di<sup>†</sup>, Luc Giraud<sup>\*</sup>,  
Xin Liang<sup>†</sup>, Nick Schenkels<sup>\*</sup>

Project-Team HiePACS

Research Report n° 9342 — May 2020 — 59 pages

**Abstract:** Large scale applications running on HPC systems often require a substantial amount of memory and can have a large computational overhead. Lossy data compression techniques can reduce the size of the data and associated communication cost, but the effect of the loss of accuracy on the numerical algorithm can be hard to predict. In this paper we examine the FGMRES algorithm, which requires the storage of a basis for the Krylov subspace and for the search space spanned by the solutions of the preconditioning systems. We show that the vectors spanning this search space can be compressed by looking at the combination of FGMRES and compression in the context of inexact Krylov subspace methods. This allows us to derive a bound on the normwise relative compression error in each iteration. We use this bound to formulate a number of different practical compression strategies, and validate and compare them through numerical experiments.

**Key-words:** flexible GMRES, inexact Krylov, compression avec perte, précision mixte.

---

<sup>\*</sup> Inria, France

<sup>†</sup> Argonne National Laboratory, IL, USA

**RESEARCH CENTRE  
BORDEAUX – SUD-OUEST**

200 avenue de la Vielle Tour  
33405 Talence Cedex

## Exploration du stockage de précision variable par des techniques de compression avec perte en algèbre linéaire numérique : une première application au GMRES flexible

**Résumé :** Les applications à grande échelle fonctionnant sur des systèmes HPC nécessitent souvent une quantité importante de mémoire et peuvent avoir une charge de calcul importante. Les techniques de compression de données avec perte peuvent réduire la taille des données et les coûts de communication associés, mais l'effet de la perte de précision sur l'algorithme numérique peut être difficile à prévoir. Dans cet article, nous examinons l'algorithme FGMRES, qui nécessite le stockage d'une base pour le sous-espace de Krylov et pour l'espace de recherche couvert par les solutions des systèmes de préconditionnement. Nous montrons que les vecteurs couvrant cet espace de recherche peuvent être comprimés en examinant la combinaison de FGMRES et de la compression dans le contexte des méthodes inexactes du sous-espace de Krylov. Cela nous permet de dériver une borne sur l'erreur de compression relative normale dans chaque itération. Nous utilisons cette limite pour formuler un certain nombre de stratégies de compression pratiques différentes, et les valider et les comparer par des expériences numériques.

**Mots-clés :** flexible GMRES, inexact Krylov, lossy compression, mixed precision.

## Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Krylov subspace methods</b>	<b>5</b>
2.1	Generalized minimal residual method . . . . .	5
2.2	Flexible generalized minimal residual method . . . . .	6
<b>3</b>	<b>Inexactness &amp; compression</b>	<b>7</b>
3.1	Inexact right preconditioning . . . . .	9
3.2	Compressed FGMRES . . . . .	10
<b>4</b>	<b>Practical compression strategies</b>	<b>11</b>
4.1	Base strategy . . . . .	12
4.2	Relaxed & double relaxed strategy . . . . .	12
4.3	Equal strategy . . . . .	12
4.4	Cast 16 & 32 bit . . . . .	13
<b>5</b>	<b>Compression details</b>	<b>13</b>
5.1	The SZ compressor . . . . .	13
5.2	Compression & memory ratios . . . . .	14
<b>6</b>	<b>Numerical experiment</b>	<b>15</b>
<b>7</b>	<b>Conclusion &amp; remarks</b>	<b>17</b>
<b>A</b>	<b>Exploiting dimension information</b>	<b>28</b>
<b>B</b>	<b>Pointwise vs. normwise</b>	<b>30</b>
<b>C</b>	<b>Detailed results for some of the matrices</b>	<b>31</b>
<b>D</b>	<b>Full tables</b>	<b>34</b>
D.1	Backtracking strategy . . . . .	34
D.2	Heuristic strategy . . . . .	34
D.3	SZ 16 & SZ 32 . . . . .	35

## 1 Introduction

Despite the arrival of the first exascale systems in the next few years, large scale applications running on high-performance computing (HPC) systems often still require a substantial amount of memory and can have a large communication overhead. This is particularly true in the context of Krylov subspace methods where a basis for the Krylov subspace may need to be stored and the calculations of dot products impose a global reduction step in the communication. Because data compression techniques reduce the size of the data, and hence the associated communication cost when the data needs to be accessed, they have been studied extensively over the last decade, and especially lossy data compression techniques have been shown to be very successful at reducing the data size in various applications [1, 7, 13, 21, 22, 30]. The obvious downside of lossy compression techniques with respect to lossless compression techniques is, of course, the loss of accuracy in the decompressed data. While these compressors allow the user to control this error, the effect on the final result of the numerical algorithm is not always easy to predict.

At the same time, in the context of numerical analysis, there has been a lot of work revolving around mixed precision algorithms [1, 2, 5, 6, 8, 9, 17]. Here, a part of the computations – often the preconditioner – is performed in single (32 bit) or half precision (16 bit), instead of the standard double precision (64 bit). The motivation for these approaches is the fact that floating point operations in single or half precision can be performed faster than those in double precision on modern computer architectures. By, for example, calculating the preconditioner in lower precision, this computationally expensive step can suddenly be performed much faster. There is, however, also a lot to be gained from reducing the communication cost and the memory requirements of numerical algorithms [1, 13, 21]. With this purpose in mind, lossy compression techniques can potentially be more flexible since they are not bound by these two data formats and can yield compression ratios much higher than 2 or 4, respectively for single and half precision. The objective of this paper is to assess such a strategy, which we will illustrate in the context of Krylov subspace methods.

We will consider the solution of linear systems  $Ax = b$ , with  $A \in \mathbb{R}^{n \times n}$  a large and sparse matrix. Krylov subspace methods remain among the most widely used methods to solve this kind of system and for non-symmetric matrices the generalized minimal residual algorithm (GMRES [23, 25, 26]) with a right preconditioner  $M \in \mathbb{R}^{n \times n}$  is often the go to method:

$$AM^{-1}u = b \quad \text{and} \quad x = M^{-1}u. \quad (1)$$

The downside of GMRES is that it requires the storage of an orthonormal basis  $V_k \in \mathbb{R}^{n \times k}$ , where  $k$  is the number of iterations. For large  $n$  and  $k$  this can be a substantial memory requirement. Furthermore, in order to allow the preconditioner to vary in each iteration, the so-called flexible generalized minimal residual method (FGMRES [16, 24, 25]) needs to be used. This is, for example, the case when another iterative method is used as a preconditioner, e.g., multigrid or GMRES itself [14, 16, 24]. As a result, FGMRES no longer uses the space spanned by the Krylov basis  $V_k$  to find the  $k$ th iteration, but rather the space spanned by the columns of  $Z_k \in \mathbb{R}^{n \times k}$ : the solutions of the preconditioning systems. This, however, means that  $Z_k$  needs to be stored as well, essentially doubling the memory required by the algorithm. While restarting strategies for these algorithms exist, they often lead to slower convergence.

In what follows, we will show that the columns of  $Z_k$  can be stored using lossy compression techniques without any loss of final accuracy in the approximation to  $x$ . We offer three different motivations as to why this approach is likely to work:

1. Typically, the preconditioning system is not solved very accurately and therefore only the first few digits are likely to be correct. If the errors introduced in the data after decompression only affect the final digits, then we will not have lost any accuracy in this step.

2. Adding compression after solving the preconditioning system can be seen as solving it in a lower precision and there are results showing the feasibility of this approach [1, 9, 17].
3. As long as  $Z_k$  has full rank, its columns can in theory be random. Although this will lead to slow convergence in practice, it does suggest that small perturbations in  $Z_k$  are unlikely to significantly effect the convergence of the method.

In order to show how lossy compression can be incorporated in FGMRES and to motivate our approach, we will look at FGMRES and compression in the context of inexact Krylov subspace methods. This is a recent development in the theory of Krylov subspace methods, which was motivated by the fact that in many applications the matrix vector product with  $A$  required in each iteration is only calculated approximately [4, 15, 29, 32].

Lossy data compression techniques fall into two main categories: prediction based and transformation based. We refer to [31] and the references therein for an overview and comparison. For our analysis, however, we only assume that the compressor can bound the normwise relative difference between the original data  $z \in \mathbb{R}^n$  and the decompressed data  $\tilde{z} \in \mathbb{R}^n$ , i.e.,

$$\frac{\|z - \tilde{z}\|}{\|z\|} \leq \zeta.$$

Here, we call  $0 < \zeta$  the maximum normwise relative compression error. From a numerical point of view this is a very natural constraint to impose, however, in order to compare our results with mixed precision approaches in half and single precision, we will also consider the maximum pointwise relative error  $0 < \varphi$ , which satisfies the inequality

$$\max_{i=1,\dots,n} \frac{|z[i] - \tilde{z}[i]|}{|z[i]|} \leq \varphi,$$

with  $z[i]$  the  $i$ th component of  $z$ .

The outline of this paper is as follows. In [section 2](#) we give a brief overview of GMRES and FGMRES. In [section 3](#) we review some results from the theory of inexact Krylov subspace methods and apply them in the context of inexact preconditioning, compression and FGMRES. We use these results in order to find a bound on the maximum normwise relative compression error  $\zeta$  that can be introduced in each iteration. In [section 4](#) we derive a number of practical compression strategies, and in [section 5](#) we discuss the SZ compressor that we will use for our numerical experiments and how the compression quality can be assessed. Finally, in [section 6](#), we perform an extensive series of numerical experiments illustrating our results.

We will refer to the exact solutions of (1) as  $u^*$  and  $x^*$ . Unless mentioned otherwise,  $\|\cdot\|$  is the standard Euclidean 2-norm  $\|\cdot\|_2$ . The largest and smallest singular values of a matrix  $A$  are denoted as  $\sigma_{max}(A)$  and  $\sigma_{min}(A)$ , respectively, and its condition number as  $\mathcal{K}(A) = \|A^{-1}\| \|A\| = \sigma_{max}(A)/\sigma_{min}(A)$ .

## 2 Krylov subspace methods

### 2.1 Generalized minimal residual method

Starting from an initial estimate  $x_0$  for  $x^*$ , GMRES constructs a series of approximations  $x_k$  in Krylov subspaces of increasing size and with decreasing residual. More specifically:

$$x_k = \arg \min_{x \in x_0 + \mathcal{K}_k(A, r_0)} \|b - Ax\|,$$



with  $r_0 = b - Ax_0$  and

$$\mathcal{K}_k(A, r_0) = \text{span}\{r_0, Ar_0, \dots, A^{k-1}r_0\}$$

the  $k$ -dimensional Krylov subspace spanned by  $A$  and  $r_0$ . In practice, a matrix  $V_k = [v_1, \dots, v_k] \in \mathbb{R}^{n \times k}$  with orthonormal columns and an upper Hessenberg matrix  $\bar{H}_k \in \mathbb{R}^{(k+1) \times k}$  are iteratively constructed using the Arnoldi procedure such that  $\text{span } V_k = \mathcal{K}_k(A, r_0)$  and

$$AV_k = V_{k+1}\bar{H}_k. \quad (2)$$

This is often referred to as the Arnoldi relation. Consequently,  $x_k = x_0 + V_k y_k$  with

$$y_k = \arg \min_{y \in \mathbb{R}^k} \|\beta e_1 - \bar{H}_k y\|,$$

where  $\beta = \|r_0\|$  and  $e_1 = (1, 0, \dots, 0)^T \in \mathbb{R}^{k+1}$ .

During the iterations, the (true) residual is given by  $r_k = b - Ax_k$ . In practice the norm of the computed residual  $\tilde{r}_k = V_{k+1}(\beta e_1 - \bar{H}_k y_k)$  will be monitored, because in exact arithmetic  $\|r_k\| = \|\tilde{r}_k\|$  and  $\|\tilde{r}_k\|$  is a free by product of GMRES. It can also be calculated without explicitly constructing  $x_k$  during each iteration – which would require an additional matrix vector product with  $V_k$  and possibly the preconditioner  $M^{-1}$ . In finite precision, however, the residual gap  $\|r_k - \tilde{r}_k\|$  might be non-zero. In our implementation we will therefore stop the GMRES iterations once the true backward error is smaller than a given tolerance  $\varepsilon > 0$ , i.e.,

$$\eta_b(x_k) = \frac{\|b - Ax_k\|}{\|b\|} = \frac{\|r_k\|}{\|b\|} \leq \varepsilon.$$

However, in order to avoid calculating  $x_k$  and the true residual  $r_k$  in each iteration, we only calculate this value if the computed backward error fulfils the same condition, i.e.,

$$\tilde{\eta}_b(x_k) = \frac{\|V_{k+1}(\beta e_1 - \bar{H}_k y_k)\|}{\|b\|} = \frac{\|\tilde{r}_k\|}{\|b\|} \leq \varepsilon.$$

An overview of right preconditioned GMRES is given in [Algorithm 1](#). Here, GMRES is applied to the linear system  $AM^{-1}u = b$ , which means that the preconditioned system is solved in each iteration ([line 4](#)) and in order to retrieve the approximation for  $x^*$  an additional multiplication with the preconditioner is required at the end ([line 15](#)). For more details about GMRES and its implementation we refer to [[23](#), [25](#), [26](#)].

## 2.2 Flexible generalized minimal residual method

When we look at [line 15](#) of [Algorithm 1](#) we see that  $x_k$  is expressed as a linear combination of the vectors  $M^{-1}v_k$ . These vectors are also calculated on [line 4](#), but since  $M$  is constant – and instead of storing them – it suffices to apply the preconditioner one additional time to the vector  $V_k y_k$  in order to calculate  $x_k$ . If the preconditioner would change in each iteration, say

$$z_k = M_k^{-1}v_k \quad (3)$$

for matrices  $M_k \in \mathbb{R}^{n \times n}$ , then

$$x_k = x_0 + Z_k y_k,$$

where  $Z_k = [z_1, \dots, z_k] \in \mathbb{R}^{n \times k}$ . In contrast to GMRES,  $Z_k$  would now need to be stored, because otherwise calculating  $x_k$  would require solving all the preconditioning systems (3) an additional time. This adaptation of GMRES leads to the flexible variant shown in [Algorithm 2](#). Obviously,

**Algorithm 1** GMRES with right preconditioning

---

```

1: input:  $A, b, x_0, \text{maxit}, \varepsilon, M$ .
2:  $r_0 = b - Ax_0, \beta = \|r_0\|$  and  $v_1 = r_0/\beta$ 
3: for  $k = 1, \dots, \text{maxit}$  do
4:    $z = M^{-1}v_k$ 
5:    $w = Az$ 
6:   for  $i = 1, \dots, k$  do
7:      $\bar{H}_{i,k} = v_i^T w$ 
8:      $w = w - \bar{H}_{i,k}v_i$ 
9:   end for
10:   $\bar{H}_{k+1,k} = \|w\|$ 
11:   $v_{k+1} = w/\bar{H}_{k+1,k}$ 
12:   $y_k = \arg \min_{y \in \mathbb{R}^k} \|\beta e_1 - \bar{H}_k y\|$ 
13:   $\tilde{r}_k = \beta e_1 - \bar{H}_k y$ 
14:  if  $\|\tilde{r}_k\| < \|b\| \varepsilon$  or  $k = \text{maxit}$  then
15:     $x_k = x_0 + M^{-1}V_k y_k$ 
16:     $r_k = b - Ax_k$ 
17:    if  $\|r_k\| < \|b\| \varepsilon$  then
18:      break
19:    end if
20:  end if
21: end for
22: output:  $x_k$ 

```

---

this approach results in increased flexibility for the preconditioner, as now, for example, an iterative method could be used as a preconditioner [14, 16, 24]. There are even some results indicating that FGMRES with a fixed preconditioner can be more stable than GMRES [3].

Note that multiplying  $AM_j^{-1}$  with any of the  $v_i$  no longer results in a vector that necessarily lies in  $\text{span} V_{k+1}$  and that neither  $\text{span}(Z_k)$  or  $\text{span}(V_k)$  are Krylov subspaces. FGMRES is therefore technically not a Krylov subspace method and the original Arnoldi relation (2) no longer holds. Instead the following relation can be proven:

$$AZ_k = V_{k+1}\bar{H}_k.$$

Furthermore, as there are no conditions on the preconditioners  $M_k^{-1}$ , as long as  $Z_k$  has full rank, the  $z_k$  can be any vector in  $\mathbb{R}^n$ . For more details we refer to [16, 24, 25, 28].

### 3 Inexactness & compression

GMRES only requires one matrix vector product with  $A$  in each iteration. In many applications, however, the matrix  $A$  is not formed explicitly, but its action on a vector is performed in a matrix free fashion. This means that instead of calculating  $Av$ , what is actually calculated is  $(A + E)v$ , for some perturbation matrix  $E \in \mathbb{R}^{n \times n}$ . This idea leads to what is referred to as “inexact Krylov subspace methods” [4, 15, 29, 32].

Again, the Arnoldi relation (2) no longer holds, but it can be shown that the following Arnoldi-like relation holds:

$$AV_k + [E_1v_1, \dots, E_kv_k] = V_{k+1}\bar{H}_k. \quad (4)$$

**Algorithm 2** FGMRES

---

```

1: input:  $A, b, x_0, \text{maxit}, \varepsilon, M$ .
2:  $r_0 = b - Ax_0, \beta = \|r_0\|$  and  $v_1 = r_0/\beta$ 
3: for  $k = 1, \dots, \text{maxit}$  do
4:    $z_k = M_k^{-1}v_k$ 
5:    $w = Az_k$ 
6:   for  $i = 1, \dots, k$  do
7:      $H_{i,k} = v_i^T w$ 
8:      $w = w - H_{i,k}v_i$ 
9:   end for
10:   $H_{k+1,k} = \|w\|$ 
11:   $v_{k+1} = w/H_{k+1,k}$ 
12:   $y_k = \arg \min_{y \in \mathbb{R}^k} \|\beta e_1 - \bar{H}_k y\|$ 
13:   $\tilde{r}_k = \beta e_1 - \bar{H}_k y$ 
14:  if  $\|\tilde{r}_k\| < \|b\| \varepsilon$  or  $k = \text{maxit}$  then
15:     $x_k = x_0 + Z_k y_k$ 
16:     $r_k = b - Ax_k$ 
17:    if  $\|r_k\| < \|b\| \varepsilon$  then
18:      break
19:    end if
20:  end if
21: end for
22: output:  $x_k$ 

```

---

It turns out that the computed residual in each iteration is given by  $\tilde{r}_k = b - \tilde{A}_k x_k$ , where  $\tilde{A}_k$  is a perturbed version of  $A$  and that

$$\tilde{A}_k V_k = V_{k+1} \bar{H}_k.$$

This means that the iterations  $x_k$  are in fact members of different Krylov subspaces, each spanned by a different matrix. Furthermore, if the size of the perturbations  $\|E_k\|$  is bounded in each iteration, it is shown in [15] that the residual gap remains small and that true residual will satisfy the stopping criterium:

**Theorem 3.1.** *Choose  $0 < \varepsilon$  and  $0 < c < 1$ . Define  $\varepsilon_c = c\varepsilon$  and  $\varepsilon_g = (1 - c)\varepsilon$ , and assume that in every inexact GMRES iteration  $k$*

$$\|E_k\| \leq \frac{c}{n} \sigma_{\min}(A) \min \left( 1, \frac{\|b\|}{\|\tilde{r}_{k-1}\|} \varepsilon_g \right). \quad (5)$$

*Then there exists an  $0 < \ell \leq n$  such that  $\|\tilde{r}_\ell\| \leq \|b\| \varepsilon_c$  and  $\|r_\ell\| \leq \|b\| \varepsilon$ .*

*Proof.* We refer the reader to [15, Theorem 2] for the full proof of this theorem.  $\square$

The key point in the proof of this theorem is that by placing a bound on  $\|E_k\|$ , the residual gap  $\|r_k - \tilde{r}_k\|$  can be bounded by  $\|b\| \varepsilon_g$ . It should also be noted that this theorem is a refinement of the results in [29] and that similar results can be found in [4, 32]. Furthermore, since this bound is inversely proportional to the computed residual, it will increase as the iterations proceed. This would imply that the initial iterations of the Krylov subspace method need to be calculated accurately, but that the final iterations can be calculated with less precision. This in contrast to other results on inexact Newton methods which indicate the opposite [11, 27].

### 3.1 Inexact right preconditioning

Suppose now that there is a fixed right preconditioner  $M$  and that the matrix vector product with  $A$  is calculated exactly, but that the solution of the preconditioned system is done inexactly. This means that [step 4 of Algorithm 2](#) is solved with residual  $p_k \neq 0$  and we can write

$$p_k = v_k - Mz \quad \Leftrightarrow \quad z = M^{-1}(v_k - p_k). \quad (6)$$

This implies that  $w = AM^{-1}(v_k - p_k)$  and the inexact Arnoldi-like relation (4) becomes

$$AM^{-1}V_k + [E_1v_1, \dots, E_kv_k] = V_{k+1}\bar{H}_k$$

with  $E_k = -AM^{-1}p_kv_k^T$ . In [15] it was shown that:

**Theorem 3.2.** *Choose  $0 < \varepsilon$  and  $0 < c < 1$ . Define  $\varepsilon_c = c\varepsilon$  and  $\varepsilon_g = (1 - c)\varepsilon$ , and assume that in every GMRES iteration  $k$  the right preconditioning system  $z = M^{-1}v_k$  is solved with residual  $p_k$ . If for all  $k$*

$$\|p_k\| \leq \frac{c}{n} \frac{1}{\mathcal{K}(AM^{-1})} \min\left(1, \frac{\|b\|}{\|\tilde{r}_{k-1}\|} \varepsilon_g\right), \quad (7)$$

*then there exists an  $0 < \ell \leq n$  such that  $\|\tilde{r}_\ell\| \leq \|b\| \varepsilon_c$  and  $\|b - AM^{-1}u_\ell\| \leq \|b\| \varepsilon$ .*

*Proof.* We refer the reader to [15, Theorem 5] for the full proof of this theorem.  $\square$

In order to retrieve  $x_\ell$  GMRES requires an additional operation of the preconditioner, introducing another residual in case of an inexact preconditioner. This is why we wrote  $\|b - AM^{-1}u_\ell\|$  in [Theorem 3.2](#) and not  $\|r_\ell\| = \|b - Ax_\ell\|$ , see [15] for more details. We, however, are interested in the link between GMRES with an inexact preconditioner and FGMRES. Suppose therefore that the preconditioner in FGMRES is an iterative method that solves the linear system  $Mz_k = v_k -$  with  $M$  fixed – up to a user defined precision. Equivalently, this could be seen as an application of GMRES with a fixed but inexact preconditioner  $M$ . Since the only real difference between the two algorithms is the fact that FGMRES stores the  $z_k$ , we will not need this final multiplication with  $M^{-1}$  in order to calculate  $x_\ell$ . Note that in practice the preconditioning system is typically not solved very accurately since, if  $M \approx A$ , this would be almost as costly as solving  $Ax = b$  directly. In the case of an iterative preconditioner, we will also make the assumption that  $M = A$ , but for generality and notational clarity will not always make this substitution.

**Theorem 3.3.** *Choose  $0 < \varepsilon$  and  $0 < c < 1$ . Define  $\varepsilon_c = c\varepsilon$  and  $\varepsilon_g = (1 - c)\varepsilon$ , and assume that in every FGMRES iteration  $k$  the right preconditioning system  $z_k = M^{-1}v_k$  is solved with residual  $p_k$ . If for all  $k$*

$$\|p_k\| \leq \frac{c}{n} \frac{1}{\mathcal{K}(AM^{-1})} \min\left(1, \frac{\|b\|}{\|\tilde{r}_{k-1}\|} \varepsilon_g\right), \quad (8)$$

*then there exists an  $0 < \ell \leq n$  such that  $\|\tilde{r}_\ell\| \|b\| \varepsilon_c$  and  $\|r_\ell\| \leq \varepsilon \|b\|$ . Note that when  $M = A$ , the term with the condition number can be dropped.*

*Proof.* This follows from [Theorem 3.2](#) and the remarks made in the previous section.  $\square$

Finally, we remark that, using (6), the FGMRES Arnoldi-like relation

$$AZ_k = V_{k+1}\bar{H}_k$$

can be written as

$$AM^{-1}V_k + [E_1v_1, \dots, E_kv_k] = V_{k+1}\bar{H}_k$$

with  $E_k = -AM^{-1}p_kv_k^T$ . This is identical to the Arnoldi-like relation for inexact preconditioned GMRES, illustrating in yet another way the similarity of both methods.

### 3.2 Compressed FGMRES

An initial idea to apply lossy compression in Krylov subspace methods might be to compress the vectors  $V_k$  in normal GMRES (or with fixed preconditioning). The problem with this approach is that these vectors are orthonormal. Each new vector  $v_{k+1}$  would be constructed using the modified Gram-Schmidt procedure applied to the decompressed version of  $V_k$ . Unfortunately, due to the loss of accuracy, these decompressed vectors are no longer orthonormal, and neither will be the new  $v_{k+1}$ . We will therefore take another approach.

As shown in the previous sections, the vectors  $z_k$  in FGMRES are the solutions of the preconditioning systems and there are results on preconditioners with lower accuracy [1, 2, 9, 17]. Furthermore, since the  $z_k$  can in theory be random – as long as  $Z_k$  is of full rank – FGMRES is likely less sensitive to small changes in these vectors. In contrast to the mixed precision approaches, however, we will perform all computations in double precision (64 bit), but store the  $z_k$  in compressed form after their calculation. An overview of this compressed FGMRES algorithm (cFGMRES) is given in Algorithm 3: it includes the compression step on line 5 and two decompression steps on lines 6 and 17. Here, the  $\tilde{z}_k$  are the vectors containing the decompressed values corresponding to the original  $z_k$ .

In order to analyse the effect of the decompression error, we will write the decompressed values  $\tilde{z}_k$  as a perturbed version of the original values  $z_k$ :

$$\tilde{z}_k = (I + F_k) z_k. \quad (9)$$

Here,  $I, F_k \in \mathbb{R}^n$  are the identity matrix and a perturbation matrix, respectively. This means that

$$\frac{\|z_k - \tilde{z}_k\|}{\|z_k\|} \leq \zeta_k, \quad (10)$$

with  $\zeta_k = \|F_k\|$  the maximum normwise relative compression error in iteration  $k$ . As mentioned before, from a numerical point of view, the only assumption we will make on the compressor is that  $\zeta_k$  can be controlled by the user.

**Theorem 3.4.** *Choose  $0 < \varepsilon$  and  $0 < c < 1$ . Define  $\varepsilon_c = c\varepsilon$  and  $\varepsilon_g = (1 - c)\varepsilon$ , and assume that in every cFGMRES iteration  $k$  the right preconditioning system  $z_k = M^{-1}v_k = A^{-1}v_k$  is solved with residual  $p_k$  and that the maximum normwise relative compression error is given by  $\eta_k > 0$ . If for all  $k$*

$$\|p_k\| + \zeta_k \|A\| \|z_k\| \leq \frac{c}{n} \min\left(1, \frac{\|b\|}{\|\tilde{r}_{k-1}\|} \varepsilon_g\right) \quad (11)$$

*then there exists an  $0 < \ell \leq n$  such that  $\|\tilde{r}_\ell\| \|b\| \varepsilon_c$  and  $\|r_\ell\| \leq \varepsilon \|b\|$ .*

*Proof.* We can interpret the compression as part of the preconditioning and write

$$\tilde{z}_k = (I + F_k) M^{-1} (v_k - p_k).$$

This means that the residual of the combined preconditioning-compression step is given by

$$v_k - M\tilde{z}_k = p_k - MF_k M^{-1} (v_k - p_k) = p_k - MF_k z_k.$$

Substituting  $M = A$  and bounding the norm of the left-hand side by

$$\|p_k\| + \zeta_k \|A\| \|z_k\|$$

the theorem now follows from Theorem 3.3. □

As before, we can also derive a new Arnoldi-like relation for cFGMRES. Because we have both an inexact preconditioner and apply compression

$$\begin{aligned} w &= A(I + F_k)M^{-1}(v_k - p_k) \\ &= AM^{-1}v_k - AM^{-1}p_k + AF_kM^{-1}(v_k - p_k). \end{aligned}$$

It follows that

$$AM^{-1}V_k + [E_1v_1, \dots, E_kv_k] = V_{k+1}\bar{H}_k,$$

with

$$E_k = (-AM^{-1}p_k + AF_kM_k^{-1}(v_k - p_k))v_k^T.$$

We already encountered the first term of  $E_k$  in the previous section as the result of the inexact preconditioner. The second term is new and is therefore due to the compression.

---

### Algorithm 3 cFGMRES

---

```

1: input:  $A, b, x_0, \text{maxit}, \varepsilon, M$ .
2:  $r_0 = b - Ax_0, \beta = \|r_0\|$  and  $v_1 = r_0/\beta$ 
3: for  $k = 1, \dots, \text{maxit}$  do
4:    $z_k = M_k^{-1}v_k$ 
5:   Compress  $z_k$ .
6:   Retrieve the decompressed vector  $\tilde{z}_k$ .
7:    $w = A\tilde{z}_k$ 
8:   for  $i = 1, \dots, k$  do
9:      $H_{i,k} = v_i^T w$ 
10:     $w = w - H_{i,k}v_i$ 
11:   end for
12:    $H_{k+1,k} = \|w\|$ 
13:    $v_{k+1} = w/H_{k+1,k}$ 
14:    $y_k = \arg \min_{y \in \mathbb{R}^k} \|\beta e_1 - \bar{H}_k y\|$ 
15:    $\tilde{r}_k = \beta e_1 - \bar{H}_k y$ 
16:   if  $\|\tilde{r}_k\| < \|b\| \varepsilon$  or  $k = \text{maxit}$  then
17:     Retrieve the decompressed columns of  $\tilde{Z}_k = [\tilde{z}_1, \dots, \tilde{z}_k]$ .
18:      $x_k = x_0 + \tilde{Z}_k y_k$ 
19:      $r_k = b - Ax_k$ 
20:     if  $\|r_k\| < \|b\| \varepsilon$  then
21:       break
22:     end if
23:   end if
24: end for
25: output:  $x$ 

```

---

## 4 Practical compression strategies

Bound (8) from [Theorem 3.3](#) and bound (11) from [Theorem 3.4](#) are both based on results from the theory of inexact Krylov subspace methods, specifically [Theorem 3.1](#). In the numerical studies performed in [\[4, 29, 32\]](#) it is, however, shown that this bound is often very restrictive and can be relaxed substantially in many applications. If we take a closer look at (8) we see that the

same holds here as well. If, for example, we start from an initial estimate  $x_0 = 0$ , then in the first iterations  $\|\tilde{r}_{k-1}\| \approx \|b\|$ ; implying that the bound is of order  $\mathcal{O}(\varepsilon)$ . In practice it is observed that the preconditioning system can be solved much less accurately. Similarly, for (11), where the bound for  $\eta_k$  can become negative if  $\|p_k\|$  is too large. These results do, however, illustrate that it is the residual of the preconditioner that is of interest when using flexible preconditioning and compression. In this section we will therefore use them as a starting point for a number of practical compression strategies, i.e., ways to set  $\zeta_k$  in each iteration.

#### 4.1 Base strategy

As stated before, in practice it is observed that  $\|p_k\|$  can be larger than what [Theorem 3.4](#) would suggest. Assuming that the FGMRES iterations without compression converge, we could ignore the preconditioning error and only try to bound the compression error using (11), i.e.,

$$\zeta_k \leq \frac{c}{n \|A\| \|z_k\|} \min \left( 1, \frac{\|b\|}{\|\tilde{r}_{k-1}\|} \varepsilon_g \right)$$

In our numerical experiment we will take  $c = 0.9$ .

#### 4.2 Relaxed & double relaxed strategy

The bound used in [Theorem 3.1](#) can be written as

$$\|E_k\| \leq \lambda_k \frac{1}{\|\tilde{r}_{k-1}\|} \varepsilon_g$$

and in [4, 29, 32] it is shown that that setting  $\lambda_k = 1$ , thus allowing larger perturbations in the matrix vector product, does not negatively impact the convergence in many cases. We will therefore do the same with the base compression strategy and relax bound (11) to find

$$\zeta_k \leq \frac{1}{\|A\| \|z_k\| \|\tilde{r}_{k-1}\|} \varepsilon_g. \quad (12)$$

If the iterations converge, we also have that  $\|\tilde{r}_{k-1}\|$  decreases to  $\varepsilon_g$ , so we can relax this bound a second time by replacing  $\varepsilon_g / \|\tilde{r}_{k-1}\|$  with 1:

$$\zeta_k \leq \frac{1}{\|A\| \|z_k\|}. \quad (13)$$

We will refer to strategy (12) and (13) as the relaxed and double relaxed strategies respectively.

#### 4.3 Equal strategy

Assuming that the FGMRES iterations without compression converge, there is a series of preconditioning errors  $\|p_k\|$  which do not prevent the algorithm from converging. Instead of using the upper bound from (11), we could relax the base strategy by using  $\|p_k\|$  as an upper bound for the maximum normwise relative compression error in each iteration, i.e.,

$$\zeta_k \|z_k\| \|A\| \leq \|p_k\| \Leftrightarrow \zeta_k \leq \frac{\|p_k\|}{\|z_k\| \|A\|}.$$

Another way to interpret this strategy is to note that [Theorem 3.4](#) suggests that it is the total perturbation from both the preconditioner and the compression that should be bounded. If the compression error in each iteration is less than or equal to the preconditioning error, then

$$\|p_k\| + \eta_k \|A\| \|z_k\| \leq 2 \|p_k\|,$$

implying that the order of magnitude of the total perturbation has remained equal to that of the FGMRES iterations without compression – which we assumed converged.

#### 4.4 Cast 16 & 32 bit

Due to the large interest in mixed precision arithmetic we will also compare the previous compression strategies with a mixed precision inspired approach: storing the  $z_k$  in either 16 bit or 32 bit precision. We will, however, perform all calculations in 64 bit, and the “decompression” step will therefore consist of casting the vector back to 64 bit. Additionally, in order to limit over- and underflow errors when casting – especially to 16 bit – we will normalize  $z_k$  before casting it and store the norm of the original data as well. After the vector is cast back to 64 bit we multiply it with its original norm in order to retrieve the “decompressed” vector  $\tilde{z}_k$ .

## 5 Compression details

### 5.1 The SZ compressor

For our numerical experiments we will use the SZ compressor. This is a prediction based compressor, meaning that it will try to predict the value of a data point based on the decompressed values of the adjacent data points in the space with dimensions corresponding to the data. The difference between the predicted and the actual value of each data point is then encoded using some quantization method. SZ specifically uses curve fitting and spline interpolation for the prediction, and error-controlled quantization and customized Huffman encoding to reduce the data size. For more details on the SZ compressor we refer to [\[12, 19, 20, 30\]](#)

From a practical point of view, SZ allows the user to control the error between the original and decompressed data using different error bounds, but we will only use two of these. Let  $z, \tilde{z} \in \mathbb{R}^n$  be the original and the decompressed data respectively, and  $0 < \chi$ . If SZ is used to compress and decompress  $z$ , then it can bound either the normwise error

$$\|z - \tilde{z}\| < \chi,$$

or the maximum pointwise relative error

$$\max_{i=1}^n \frac{|z[i] - \tilde{z}[i]|}{|z[i]|} < \chi.$$

Since SZ is a prediction based lossy compressor, the level of compression that can be achieved will be different for different data sets, even if the same error bound  $\chi$  is used. Bigger error bounds will typically lead to more compression, but also when the data is in some “smooth” or “regular”, then predictions for the data values made by SZ will be more accurate, and higher compression ratios can be expected.

The compression strategies discussed in [section 4](#) are based on the normwise relative error [\(10\)](#). This means that we will pass the value  $\chi_k = \zeta_k \|z_k\|$  to SZ. However, when casting to 16 or



32 bit, the compression error will be pointwise, and it is easy to see that this is stricter than the normwise control:

$$\forall i \in \{1, \dots, n\} : |z[i] - \tilde{z}[i]| < \zeta |z[i]| \quad \Rightarrow \quad \|z - \tilde{z}\| < \zeta \|z\|.$$

Furthermore, controlling the normwise relative error can lead to very high pointwise relative errors. Take for example  $z = (1, \dots, 1)^T \in \mathbb{R}^n$  and  $\tilde{z} = (\alpha, 1, \dots, 1)^T \in \mathbb{R}^n$ . The maximum pointwise relative error will be  $|1 - \alpha|$ , but the normwise relative error will be  $|1 - \alpha|/\sqrt{n}$ . For large  $n$  the difference for this example could be very big.

Finally, we note that in [line 7](#) of [Algorithm 3](#) we could also use the exact  $z_k$  instead of the decompressed version  $\tilde{z}_k$ . This way, we would only need to perform the decompression at the end when  $x_k$  is calculated – [line 18](#). This, however, creates a discrepancy between the vectors that are used to compute the Krylov vectors  $V_k$  and the solution  $x_k$ . This translates to the coefficients  $y_k$  and we would deviate even further from the original FGMRES algorithm. While this approach can still converge, we observed in numerical experiments not reported here, that the residual gap could become very big, leading to a less stable version of the algorithm. We will therefore limit ourselves to the version with two decompression steps as shown in [Algorithm 3](#).

## 5.2 Compression & memory ratios

In order to assess the quality of the different compression strategies we look at the memory required by FGMRES and cFGMRES to store the  $z_k$  and  $\tilde{z}_k$  respectively. Let  $\bar{\cdot}$  denote the compressed data object and  $\#(\cdot)$  the memory used by an object. Since  $\#(z_k)$  is equal for all  $k$ , we will simply write  $\#(z)$ . We now define the compression ratio of  $z_k$  in iteration  $k$  as

$$\rho_k = \frac{\#(z_k)}{\#(\tilde{z}_k)}.$$

Note that the memory used by  $\tilde{z}_k$  can vary because the compression ratio depends on  $z_k$  itself and on the bound for the pointwise relative error – which will vary in each iteration. If FGMRES needs  $\ell_{ref}$  iterations to converge and cFGMRES  $\ell$  iterations then we define the compression ratio associated with  $Z_\ell = [z_1, \dots, z_\ell]$  as

$$\rho = \frac{\sum_{k=1}^{\ell_{ref}} \#(z_k)}{\sum_{k=1}^{\ell} \#(\tilde{z}_k)} = \frac{\ell_{ref} \cdot \#(z)}{\sum_{k=1}^{\ell} \frac{\#(z)}{\rho_k}} = \frac{\ell_{ref}}{\sum_{k=1}^{\ell} \frac{1}{\rho_k}}. \quad (14)$$

The compression ratio gives us an easy way to assess the overall efficiency of the compression, taking into account the difference in the number of iterations. We might, for example, have a high compression ratio in each iteration, but if we need many extra iterations to converge, we still may find  $\rho < 1$ .

In order to estimate how much memory we gain with respect to FGMRES we also define the memory ratio:

$$\begin{aligned} \mu &= \frac{\sum_{k=1}^{\ell_{ref}} \#(v_k) + \#(z_k)}{\sum_{k=1}^{\ell} \#(v_k) + \#(\tilde{z}_k)} = \frac{\ell_{ref} \cdot (\#(v) + \#(z))}{\ell \cdot \#(v) + \sum_{k=1}^{\ell} \#(\tilde{z}_k)} \\ &= \frac{2\ell_{ref} \cdot \#(z)}{\ell \cdot \#(z) + \sum_{k=1}^{\ell} \frac{\#(z)}{\rho_k}} \\ &= \frac{2\ell_{ref}}{\ell + \sum_{k=1}^{\ell} \frac{1}{\rho_k}}. \end{aligned} \quad (15)$$

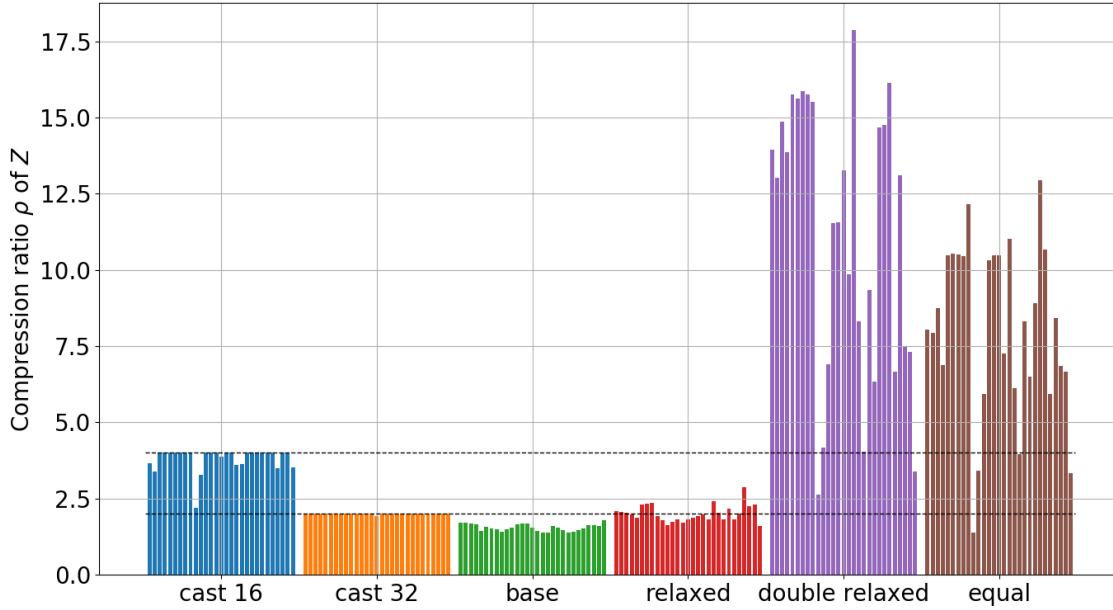


Figure 1: The compression ratio  $\rho$  of  $Z$  grouped per compression strategy. Each bar per strategy corresponds to a different matrix according to its id in Table 1. Horizontal reference lines are added at  $\rho = 2$  and 4.

This takes into account the storage required for both the  $v_k$  and  $z_k$  (or  $\bar{z}_k$ ) and uses the fact that  $\#(v_k) = \#(v) = \#(z)$  for all  $k$ . Obviously, higher compression ratios in each iteration will lead to a higher total compression and memory ratio. The latter will, however, penalize extra iterations a lot more than the former since it takes into account the fact that the extra  $v_k$  need to be stored as well – without compression. Note that we can write

$$\mu(\rho) = \frac{2\ell_{ref}\rho}{\rho\ell + \ell_{ref}} \quad \Rightarrow \quad \lim_{\rho \rightarrow +\infty} \mu(\rho) = 2\frac{\ell_{ref}}{\ell}.$$

While it is possible that  $\ell \leq \ell_{ref}$ , we will see in our numerical experiments that the opposite is usually true. This implies that the memory ratio is bounded by 2, which is not surprising, since even with very high compression rates cFGMRES still needs to store the  $z_k$ . Also note that when the compression is done by casting the  $z_k$  to 16 bit and  $\ell = \ell_{ref}$ , then  $\rho = 4$  and  $\mu = 1.6$ . Similarly, for casting to 32 bit we find  $\rho = 2$  and  $\mu = 4/3$ .

## 6 Numerical experiment

We consider a number of linear systems  $Ax = b$  of size  $n \times n$ , where the solution  $x^*$  is a vector with random uniform entries drawn from  $[-1, 1]$ , and solve them using cFGMRES with the different compression strategies from section 4. As a preconditioner we will use normal GMRES with  $\varepsilon = 1e-1$  and `maxit` = 5. For cFGMRES itself we take  $\varepsilon = 1e-10$  and `maxit` equal to twice the number of iterations required by FGMRES to converge – more iterations would negate any memory gain that compression could have. For practical reasons we also limit the values that can be passed to the SZ compressor by imposing the following restrictions:  $1e-18 \leq \chi_k \leq 1$ .

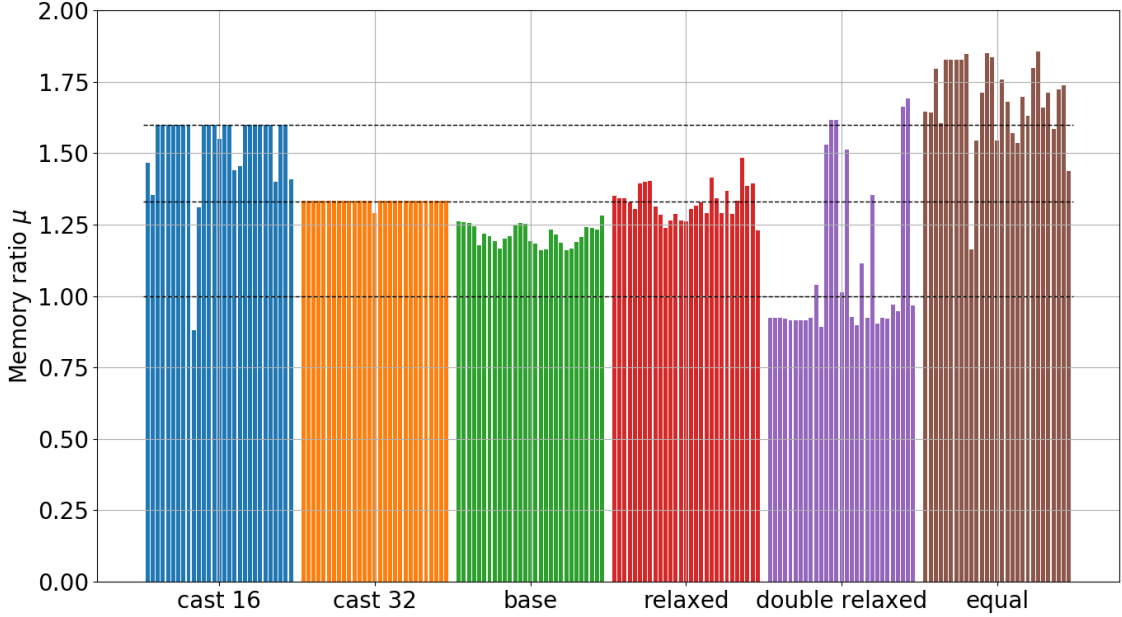


Figure 2: The memory ratio  $\mu$  grouped per compression strategy. Each bar per strategy corresponds to a different matrix according to its id in Table 1. Horizontal reference lines are added at  $\mu = 1$ , 1.33, and 1.6.

In order to remove any effect of the matrices being badly scaled, we also scale the matrices using algorithm 1 from [18]. This algorithm scales the rows and columns of a matrix  $A$  so that each has  $\infty$ -norm equal to 1. It is well known that such scaling can be computationally beneficial, but it can also be used in a mixed-precision context, as explored in [17]. Table 1 lists some basic information on the matrices we use, as well as the number of iterations FGMRES-GMRES requires to solve either the original or the scaled system.

In tables 2–7 we show the results of the different compression strategies. Each table shows the number of iterations performed and its value relatively to the number of iterations required when no compression is used, i.e., the number of iterations from Table 1. Furthermore, we list the backward error  $\eta_b$ , the compression and memory ratios  $\rho$  and  $\mu$ , and the minimum and maximum value of the normwise relative and pointwise error during the iterations  $\zeta_k$  and  $\varphi_k$ . Note that for the casting strategies we have no explicit control over these values, so we calculate them after decompression. For the compression strategies using SZ, we do control  $\zeta_k$ , but not  $\varphi_k$ , since we use the normwise error bound control, so again, we calculate this second value after the decompression. The a posteriori calculation of these values is indicated by  $\tilde{\cdot}$ . Figures 1 and 2 show a summary of the compression ratios  $\rho$  of  $Z$  and memory ratios  $\mu$  for the different compression strategies.

As can be seen in figures 1 and 2, the cast 16 compression strategy achieves its maximum compression ratio  $\rho = 4$  and memory  $\mu = 1.6$  for many of the matrices. Perhaps even more remarkable is the fact that this strategy converges at all given that the maximum pointwise relative error  $\varphi$  is almost always 1, indicating that in each iteration, in at least one component of  $z_k$ , all information was lost, see Table 2. This is not the case for the cast 32 strategy, see Table 3, but here the compression ratios are lower ( $\rho = 4$  and  $\mu = 1.33$ ). The results of this strategy are, however, more stable with respect to the theoretically expected outcome.

As expected, we see that the base strategy does not achieve high compression ratios, see Table

4. This is coherent with the observations for inexact Krylov subspace methods, and manifests itself here in the very small values of  $\zeta_k$ . It is impossible to expect high compression ratios for such tight error bounds. Relaxing the bound only slightly improves the compression ratios, see [Table 5](#). This is because while  $\zeta_k$  can now become larger, this only occurs in the final iterations and is not enough to account for its lower values in the first iterations. The double relaxed strategy does result in large values of  $\zeta_k$  in every iteration, see [Table 6](#), which translates into high compression ratios, see [Figure 1](#). However, because so much information is lost due to the high allowed compression error, this strategy does not manage to converge within twice the number of normal iterations for almost all the matrices. This means that the memory ratio is smaller than 1, and we do not gain anything, see [Figure 2](#). By contrast, the equal compression strategy also yields high values for  $\zeta_k$ , but is much more stable and converges for all matrices. As a result, both the compression and memory ratios are very high, outperforming the cast 16 compression strategy in almost all cases.

## 7 Conclusion & remarks

In this paper we studied how lossy compression can be used to compress the FGMRES search space  $Z_k$  in order to reduce the memory used by the algorithm. For this, we derived a theoretical framework by linking the compression with inexact Krylov subspace methods and inexact preconditioning, resulting in [theorems 3.3](#) and [3.4](#). We also formulated a number of different compression strategies, which we tested through a series of numerical experiments. Here we observed that a simple cast 16 strategy can be effective in many cases, but that much higher compression ratios can be achieved using lossy compressors. This illustrates that it might be interesting to look further than classical framework of double, single, and half precisions.

For these experiments we used the SZ compressor, but our results are independent from this choice. The only assumption we make on the compressor is that we can bound the normwise relative error between the original and the decompressed data. In experiments not reported here, we replaced the compression/decompression step with simply adding a perturbation of a given size to the vector and saw similar convergence behaviour. We therefore expect to see the same results when a different compressor is used. The precise compression level will, however, vary depending on which compressor is used for which linear system.

The result presented so far are a summery of the principle results from various numerical experiments we performed. On top of the results presented here we performed other experiments using other compression strategies, performed all experiments with and without the row- and column scaling of the matrices, and studied the effect of passing information about the dimensions of the linear system, i.e., does it represent a 2D, 3D, . . . , problem, to the SZ compressor. We refer to the attached appendices for the full results and a discussion thereof.

## Acknowledgements

Experiments presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see <https://www.plafrim.fr/>).

## References

- [1] H. ANZT, J. DONGARRA, G. FLEGAR, N. J. HIGHAM, AND E. S. QUINTANA-ORTÍ,

- Adaptive precision in block-jacobi preconditioning for iterative sparse linear system solvers*, Concurrency and Computation: Practice and Experience, 31 (2019), p. e4460.
- [2] M. ARIOLI AND I. S. DUFF, *Using FGMRES to obtain backward stability in mixed precision*, Electronic Transactions on Numerical Analysis, 33 (2009), pp. 31–44.
- [3] M. ARIOLI, I. S. DUFF, S. GRATTON, AND S. PRALET, *A note on GMRES preconditioned by a perturbed  $LDL^T$  decomposition with static pivoting*, SIAM Journal on Scientific Computing, 29 (2007), pp. 2024–2044.
- [4] A. BOURAS AND V. FRAYSSÉ, *Inexact matrix-vector products in Krylov methods for solving linear systems: a relaxation strategy*, SIAM Journal on Matrix Analysis and Applications, 26 (2005), pp. 660–678.
- [5] A. BUTTARI, J. DONGARRA, J. KURZAK, P. LUSZCZEK, AND S. TOMOV, *Using mixed precision for sparse matrix computations to enhance the performance while achieving 64-bit accuracy*, ACM Transactions on Mathematical Software (TOMS), 34 (2008), p. 17.
- [6] A. BUTTARI, J. DONGARRA, J. LANGOU, J. LANGOU, P. LUSZCZEK, AND J. KURZAK, *Mixed precision iterative refinement techniques for the solution of dense linear systems*, The International Journal of High Performance Computing Applications, 21 (2007), pp. 457–466.
- [7] J. CALHOUN, F. CAPPELLO, L. N. OLSON, M. SNIR, AND W. D. GROPP, *Exploring the feasibility of lossy compression for PDE simulations*, The International Journal of High Performance Computing Applications, 33 (2019), pp. 397–410.
- [8] E. CARSON AND N. J. HIGHAM, *A new analysis of iterative refinement and its application to accurate solution of ill-conditioned sparse linear systems*, SIAM Journal on Scientific Computing, 39 (2017), pp. A2834–A2856.
- [9] E. CARSON AND N. J. HIGHAM, *Accelerating the solution of linear systems by iterative refinement in three precisions*, SIAM Journal on Scientific Computing, 40 (2018), pp. A817–A847.
- [10] T. A. DAVIS AND Y. HU, *The University of Florida sparse matrix collection*, ACM Transactions on Mathematical Software (TOMS), 38 (2011), p. 1.
- [11] R. S. DEMBO, S. C. EISENSTAT, AND T. STEIHAUG, *Inexact Newton methods*, SIAM Journal on Numerical analysis, 19 (1982), pp. 400–408.
- [12] S. DI AND F. CAPPELLO, *Fast error-bounded lossy hpc data compression with SZ*, in 2016 IEEE international parallel and distributed processing symposium (ipdps), IEEE, 2016, pp. 730–739.
- [13] S. DI AND F. CAPPELLO, *Optimization of error-bounded lossy compression for hard-to-compress HPC data*, IEEE transactions on parallel and distributed systems, 29 (2017), pp. 129–143.
- [14] S. GAZZOLA AND M. S. LANDMAN, *Flexible GMRES for total variation regularization*, BIT Numerical Mathematics, 59 (2019), pp. 721–746.
- [15] L. GIRAUD, S. GRATTON, AND J. LANGOU, *Convergence in backward error of relaxed GMRES*, SIAM Journal on Scientific Computing, 29 (2007), pp. 710–728.

- 
- [16] L. GIRAUD, S. GRATTON, X. PINEL, AND X. VASSEUR, *Flexible GMRES with deflated restarting*, SIAM Journal on Scientific Computing, 32 (2010), pp. 1858–1878.
- [17] N. J. HIGHAM, S. PRANESH, AND M. ZOUNON, *Squeezing a matrix into half precision, with an application to solving linear systems*, SIAM Journal on Scientific Computing, 41 (2019), pp. A2536–A2551.
- [18] P. A. KNIGHT, D. RUIZ, AND B. UÇAR, *A symmetry preserving algorithm for matrix scaling*, SIAM journal on Matrix Analysis and Applications, 35 (2014), pp. 931–955.
- [19] X. LIANG, S. DI, D. TAO, Z. CHEN, AND F. CAPPELLO, *An efficient transformation scheme for lossy data compression with point-wise relative error bound*, in 2018 IEEE International Conference on Cluster Computing (CLUSTER), IEEE, 2018, pp. 179–189.
- [20] X. LIANG, S. DI, D. TAO, S. LI, S. LI, H. GUO, Z. CHEN, AND F. CAPPELLO, *Error-controlled lossy compression optimized for high compression ratios of scientific datasets*, in 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 438–447.
- [21] P. LINDSTROM, *Fixed-rate compressed floating-point arrays*, IEEE transactions on visualization and computer graphics, 20 (2014), pp. 2674–2683.
- [22] P. LINDSTROM AND M. ISENBURG, *Fast and efficient compression of floating-point data*, IEEE transactions on visualization and computer graphics, 12 (2006), pp. 1245–1250.
- [23] C. C. PAIGE, M. ROZLOZNÍK, AND Z. STRAKOS, *Modified Gram-Schmidt (mgs), least squares, and backward stability of MGS-GMRES*, SIAM Journal on Matrix Analysis and Applications, 28 (2006), pp. 264–284.
- [24] Y. SAAD, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM Journal on Scientific Computing, 14 (1993), pp. 461–469.
- [25] Y. SAAD, *Iterative methods for sparse linear systems*, vol. 82, SIAM, 2003.
- [26] Y. SAAD AND M. H. SCHULTZ, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on scientific and statistical computing, 7 (1986), pp. 856–869.
- [27] A. H. SHERMAN, *On Newton-iterative methods for the solution of systems of nonlinear equations*, SIAM Journal on Numerical Analysis, 15 (1978), pp. 755–771.
- [28] V. SIMONCINI AND D. B. SZYLD, *Flexible inner-outer Krylov subspace methods*, SIAM Journal on Numerical Analysis, 40 (2002), pp. 2219–2239.
- [29] V. SIMONCINI AND D. B. SZYLD, *Theory of inexact Krylov subspace methods and applications to scientific computing*, SIAM Journal on Scientific Computing, 25 (2003), pp. 454–477.
- [30] D. TAO, S. DI, Z. CHEN, AND F. CAPPELLO, *Significantly improving lossy compression for scientific data sets based on multidimensional prediction and error-controlled quantization*, in 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS), IEEE, 2017, pp. 1129–1139.
- [31] D. TAO, S. DI, X. LIANG, Z. CHEN, AND F. CAPPELLO, *Optimizing lossy compression rate-distortion from automatic online selection between sz and zfp*, IEEE Transactions on Parallel and Distributed Systems, 30 (2019), pp. 1857–1871.

- [32] J. VAN DEN ESHOF AND G. L. G. SLEIJPEN, *Inexact Krylov subspace methods for linear systems*, SIAM Journal on Matrix Analysis and Applications, 26 (2004), pp. 125–153.

	name	n	nnz	$\ A\ $	original		scaled	
					iter	$\eta_b$	iter	$\eta_b$
1	atmosmodd	1270432	8814880	1.92e+05	11	1.38e-11	11	1.38e-11
2	atmosmodj	1270432	8814880	1.92e+05	11	8.43e-11	11	8.43e-11
3	atmosmodl	1489752	10319760	6.20e+05	10	1.37e-11	10	1.37e-11
4	atmosmodm	1489752	10319760	6.39e+06	10	1.15e-11	10	1.15e-11
5	cage12	130228	2032536	1.02e+00	8	3.46e-11	8	5.42e-12
6	cage13	445315	7479343	1.02e+00	8	5.38e-11	8	3.07e-12
7	cage14	1505785	27130349	1.02e+00	8	5.28e-11	8	7.68e-12
8	cage15	5154859	99199551	1.02e+00	8	9.45e-11	8	6.46e-12
9	crashbasis	160000	1750416	6.54e+02	10	3.15e-11	10	3.82e-11
10	dc1	116835	766396	5.70e+04	139	9.66e-11	11	2.12e-11
11	dc2	116835	766396	5.84e+04	89	8.80e-11	9	2.13e-11
12	dc3	116835	766396	6.25e+04	131	9.71e-11	31	1.28e-11
13	Goodwin_095	100037	3226066	1.05e+00	245	9.72e-11	120	9.93e-11
14	Goodwin_127	178437	5778545	1.05e+00	169	9.66e-11	159	9.83e-11
15	hcircuit	105676	513072	8.63e+01	215	9.58e-11	30	7.28e-11
16	language	399130	1216334	2.91e+01	9	3.40e-11	9	3.34e-11
17	majorbasis	160000	1750416	1.45e+02	10	4.67e-11	10	2.36e-11
18	memchip	2707524	13343948	5.00e+02	68	8.18e-11	9	4.69e-11
19	ML_Laplace	377002	27582698	2.92e+07	53	8.50e-11	20	4.38e-11
20	rajat31	4690002	20316253	1.25e+04	26	5.26e-11	17	6.19e-11
21	ss	1652680	34753577	6.54e+00	10	5.62e-11	28	9.28e-11
22	ss1	205282	845089	2.17e+00	7	2.74e-11	7	2.73e-11
23	stomach	213360	3021648	2.21e+00	10	4.00e-11	10	2.73e-11
24	torso2	115967	1033473	8.06e+00	10	2.60e-11	9	8.87e-11
25	trans5	116835	749800	1.13e+04	417	9.56e-11	11	1.20e-11
26	Transport	1602111	23487281	1.00e+00	34	7.55e-11	28	9.25e-11
27	vas_stokes_1M	1090664	34767207	8.85e+00	76	8.57e-11	72	7.93e-11
28	vas_stokes_2M	2146677	65129037	8.19e+00	72	5.77e-11	65	8.84e-11
29	xenon2	157464	3866688	5.29e+28	22	7.87e-11	22	8.94e-11

Table 1: Matrices taken from the SuiteSparse Matrix Collection [10] that are used in the experiments: we list the size (n), the number of non-zeros (nnz), and the norm ( $\|A\|$ ). Scaling the matrix can greatly reduce the number of iterations required by FGMRES-GRMRES to converge.



	name	iter	rel iter	$\eta_b$	$\rho$	$\mu$	min $\tilde{\zeta}_k$	max $\tilde{\zeta}_k$	min $\tilde{\varphi}_k$	max $\tilde{\varphi}_k$
1	atmosmodd	12	1.09	2.20e-11	3.67	1.47	1.99e-04	2.11e-04	1	1
2	atmosmodj	13	1.18	1.46e-11	3.38	1.35	2.06e-04	2.12e-04	1	1
3	atmosmodl	10	1.00	5.44e-11	4.00	1.60	2.04e-04	2.12e-04	1	1
4	atmosmodm	10	1.00	2.18e-11	4.00	1.60	2.06e-04	2.08e-04	1	1
5	cage12	8	1.00	5.44e-12	4.00	1.60	2.03e-04	2.16e-04	7.87e-02	1
6	cage13	8	1.00	3.08e-12	4.00	1.60	2.05e-04	2.17e-04	1	1
7	cage14	8	1.00	7.65e-12	4.00	1.60	1.99e-04	2.21e-04	1	1
8	cage15	8	1.00	6.46e-12	4.00	1.60	2.04e-04	2.12e-04	1	1
9	crashbasis	10	1.00	3.77e-11	4.00	1.60	2.06e-04	2.08e-04	1	1
10	dc1	20	1.82	1.34e-11	2.20	0.88	1.93e-04	2.35e-04	1	1
11	dc2	11	1.22	7.48e-12	3.27	1.31	1.92e-04	2.56e-04	1	1
12	dc3	31	1.00	1.87e-11	4.00	1.60	1.93e-04	2.26e-04	1.77e-01	1
13	Goodwin_095	120	1.00	1.00e-10	4.00	1.60	2.03e-04	2.11e-04	3.15e-01	1
14	Goodwin_127	159	1.00	9.80e-11	4.00	1.60	2.04e-04	2.12e-04	3.19e-01	1
15	hcircuit	31	1.03	4.47e-11	3.87	1.55	1.84e-04	2.33e-04	1	1
16	language	9	1.00	3.34e-11	4.00	1.60	1.90e-04	2.23e-04	1	1
17	majorbasis	10	1.00	2.36e-11	4.00	1.60	1.98e-04	2.12e-04	1	1
18	memchip	10	1.11	2.26e-11	3.60	1.44	1.17e-04	2.35e-04	1	1
19	ML_Laplace	22	1.10	6.36e-11	3.64	1.45	1.95e-04	2.15e-04	1	1
20	rajat31	17	1.00	9.66e-11	4.00	1.60	2.09e-04	2.10e-04	1	1
21	ss	28	1.00	9.07e-11	4.00	1.60	2.06e-04	2.12e-04	1	1
22	ss1	7	1.00	2.73e-11	4.00	1.60	1.87e-04	2.11e-04	1	1
23	stomach	10	1.00	2.73e-11	4.00	1.60	1.61e-04	2.58e-04	1	1
24	torso2	9	1.00	8.88e-11	4.00	1.60	1.91e-04	2.17e-04	1	1
25	trans5	11	1.00	7.28e-11	4.00	1.60	1.85e-04	2.20e-04	1	1
26	Transport	32	1.14	8.00e-11	3.50	1.40	2.05e-04	2.13e-04	1	1
27	vas_stokes_1M	72	1.00	8.54e-11	4.00	1.60	2.00e-04	2.12e-04	1	1
28	vas_stokes_2M	65	1.00	7.47e-11	4.00	1.60	2.03e-04	2.14e-04	1	1
29	xenon2	25	1.14	7.15e-11	3.52	1.41	2.04e-04	2.11e-04	1	1

Table 2: Results from cFGMRES-GMRES with the cast 16 compression strategy. In almost all cases we find a compression ratio  $\rho$  for  $Z$  equal to 4.

	name	iter	rel iter	$\eta_b$	$\rho$	$\mu$	$\min \tilde{\zeta}_k$	$\max \tilde{\zeta}_k$	$\min \tilde{\varphi}_k$	$\max \tilde{\varphi}_k$
1	atmosmodd	11	1.00	1.38e-11	2.00	1.33	2.46e-08	2.60e-08	5.95e-08	5.96e-08
2	atmosmodj	11	1.00	8.43e-11	2.00	1.33	2.51e-08	2.61e-08	5.95e-08	5.96e-08
3	atmosmodl	10	1.00	1.37e-11	2.00	1.33	2.49e-08	2.58e-08	5.95e-08	5.96e-08
4	atmosmodm	10	1.00	1.15e-11	2.00	1.33	2.51e-08	2.54e-08	5.95e-08	5.96e-08
5	cage12	8	1.00	5.42e-12	2.00	1.33	2.47e-08	2.64e-08	5.94e-08	5.96e-08
6	cage13	8	1.00	3.07e-12	2.00	1.33	2.43e-08	2.65e-08	5.93e-08	5.96e-08
7	cage14	8	1.00	7.68e-12	2.00	1.33	2.50e-08	2.61e-08	5.95e-08	5.96e-08
8	cage15	8	1.00	6.46e-12	2.00	1.33	2.50e-08	2.56e-08	5.95e-08	5.96e-08
9	crashbasis	10	1.00	3.82e-11	2.00	1.33	2.51e-08	2.53e-08	5.94e-08	5.95e-08
10	dc1	11	1.00	2.09e-11	2.00	1.33	2.48e-08	2.74e-08	5.92e-08	5.96e-08
11	dc2	9	1.00	2.15e-11	2.00	1.33	2.31e-08	2.65e-08	5.93e-08	5.95e-08
12	dc3	31	1.00	8.51e-12	2.00	1.33	2.31e-08	3.32e-08	5.90e-08	5.96e-08
13	Goodwin_095	120	1.00	9.93e-11	2.00	1.33	2.48e-08	2.57e-08	5.91e-08	5.96e-08
14	Goodwin_127	159	1.00	9.83e-11	2.00	1.33	2.49e-08	2.59e-08	5.93e-08	5.96e-08
15	hcircuit	31	1.03	5.54e-11	1.94	1.29	2.07e-08	2.66e-08	5.93e-08	5.96e-08
16	language	9	1.00	3.34e-11	2.00	1.33	2.24e-08	2.63e-08	5.94e-08	5.96e-08
17	majorbasis	10	1.00	2.36e-11	2.00	1.33	2.45e-08	2.60e-08	5.94e-08	5.96e-08
18	memchip	9	1.00	4.69e-11	2.00	1.33	2.21e-08	2.56e-08	5.96e-08	5.96e-08
19	ML_Laplace	20	1.00	4.38e-11	2.00	1.33	2.40e-08	2.64e-08	5.94e-08	5.96e-08
20	rajat31	17	1.00	6.19e-11	2.00	1.33	2.53e-08	2.55e-08	5.95e-08	5.96e-08
21	ss	28	1.00	9.28e-11	2.00	1.33	2.51e-08	2.59e-08	5.95e-08	5.96e-08
22	ss1	7	1.00	2.73e-11	2.00	1.33	2.29e-08	2.57e-08	5.93e-08	5.95e-08
23	stomach	10	1.00	2.73e-11	2.00	1.33	1.86e-08	2.95e-08	5.94e-08	5.96e-08
24	torso2	9	1.00	8.87e-11	2.00	1.33	2.45e-08	2.67e-08	5.93e-08	5.95e-08
25	trans5	11	1.00	1.20e-11	2.00	1.33	1.85e-08	2.72e-08	5.93e-08	5.95e-08
26	Transpuit	28	1.00	9.25e-11	2.00	1.33	2.51e-08	2.56e-08	5.95e-08	5.96e-08
27	vas_stokes_1M	72	1.00	8.24e-11	2.00	1.33	2.43e-08	2.58e-08	5.95e-08	5.96e-08
28	vas_stokes_2M	65	1.00	8.73e-11	2.00	1.33	2.49e-08	2.60e-08	5.95e-08	5.96e-08
29	xenon2	22	1.00	8.94e-11	2.00	1.33	2.50e-08	2.55e-08	5.93e-08	5.96e-08

Table 3: Results from cFGMRES-GMRES with the cast 32 compression strategy. The number of iterations required to converge is the same as when no compression is used, but, in contrast to the cast 16 strategy, the compression ratio  $\rho$  for  $Z$  is now limited to 2.

	name	iter	rel iter	$\eta_b$	$\rho$	$\mu$	min $\zeta_k$	max $\zeta_k$	min $\tilde{\varphi}_k$	max $\tilde{\varphi}_k$
1	atmosmodd	11	1.00	1.38e-11	1.71	1.26	9.62e-13	7.35e-10	9.46e-09	4.77e-02
2	atmosmodj	11	1.00	8.43e-11	1.69	1.26	9.62e-13	4.99e-10	1.26e-08	8.05e-03
3	atmosmodl	10	1.00	1.37e-11	1.68	1.25	1.53e-12	1.46e-09	7.92e-09	6.87e-03
4	atmosmodm	10	1.00	1.15e-11	1.65	1.24	6.70e-12	2.74e-09	4.31e-09	4.34e-04
5	cage12	8	1.00	5.42e-12	1.43	1.18	6.10e-15	3.11e-06	0	2.19e+02
6	cage13	8	1.00	3.07e-12	1.56	1.22	3.26e-15	8.57e-07	7.67e-12	1.05e+00
7	cage14	8	1.00	7.68e-12	1.53	1.21	1.81e-15	2.30e-07	1.06e-10	1.55e-01
8	cage15	8	1.00	6.46e-12	1.48	1.19	9.70e-16	6.72e-08	3.54e-11	1.98e-01
9	crashbasis	10	1.00	3.82e-11	1.40	1.17	9.55e-14	6.34e-07	0	2.46e-01
10	dc1	11	1.00	2.12e-11	1.50	1.20	1.69e-14	1.32e-10	2.80e-09	2.00e-03
11	dc2	9	1.00	2.13e-11	1.53	1.21	1.88e-14	2.85e-09	8.90e-10	1.78e-04
12	dc3	31	1.00	2.26e-11	1.66	1.25	1.89e-14	2.42e-08	5.14e-10	8.22e-02
13	Goodwin_095	120	1.00	9.93e-11	1.68	1.25	5.84e-15	1.60e-07	6.16e-13	1.72e+03
14	Goodwin_127	159	1.00	9.83e-11	1.68	1.25	4.40e-15	8.97e-08	1.36e-13	3.59e+01
15	hcircuit	31	1.03	4.66e-11	1.55	1.19	9.83e-15	2.72e-08	0	1.19e+01
16	language	9	1.00	3.34e-11	1.45	1.18	6.35e-16	5.60e-08	0	1.61e-02
17	majorbasis	10	1.00	2.36e-11	1.38	1.16	4.77e-14	5.47e-07	0	5.39e-02
18	memchip	9	1.00	4.69e-11	1.39	1.16	1.10e-15	5.32e-10	0	2.85e-01
19	ML_Laplace	20	1.00	4.38e-11	1.60	1.23	8.70e-12	9.15e-10	2.92e-09	1.82e+01
20	rajat31	17	1.00	6.19e-11	1.55	1.21	7.73e-15	1.21e-10	4.28e-08	8.06e+00
21	ss	28	1.00	9.28e-11	1.46	1.19	1.19e-15	1.76e-09	3.24e-11	2.05e+00
22	ss1	7	1.00	2.73e-11	1.39	1.16	5.13e-15	8.64e-07	0	1.02e-01
23	stomach	10	1.00	2.73e-11	1.40	1.17	4.82e-15	7.43e-07	0	2.46e-02
24	torso2	9	1.00	8.87e-11	1.47	1.19	1.47e-14	1.83e-06	0	7.55e-02
25	trans5	11	1.00	1.20e-11	1.52	1.21	2.09e-14	8.75e-09	0	7.30e-01
26	Transport	28	1.00	9.25e-11	1.64	1.24	1.16e-15	2.15e-10	7.96e-13	2.39e-01
27	vas_stokes_1M	72	1.00	8.11e-11	1.62	1.24	1.37e-15	3.83e-09	5.18e-11	1.17e+00
28	vas_stokes_2M	65	1.00	8.84e-11	1.61	1.23	9.04e-16	1.34e-09	5.84e-11	3.03e+00
29	xenon2	22	1.00	8.94e-11	1.78	1.28	1.62e-09	3.05e-06	2.66e-03	6.06e-01

Table 4: Results from cFGMRES-GMRES with the base compression strategy. Because the theory results in very small values for  $\zeta_k$  it is impossible for the compressor to achieve high compression ratios.

	name	iter	rel iter	$\eta_b$	$\rho$	$\mu$	min $\zeta_k$	max $\zeta_k$	min $\tilde{\varphi}_k$	max $\tilde{\varphi}_k$
1	atmosmodd	11	1.00	1.49e-11	2.07	1.35	6.11e-12	4.04e-05	1.05e-07	9.80e+02
2	atmosmodj	11	1.00	8.45e-11	2.04	1.34	6.11e-12	8.18e-06	1.51e-07	6.61e+01
3	atmosmodl	10	1.00	1.48e-11	2.04	1.34	5.95e-12	8.64e-05	2.83e-08	4.33e+02
4	atmosmodm	10	1.00	1.31e-11	1.97	1.33	7.33e-12	2.23e-04	9.81e-09	1.83e+04
5	cage12	8	1.00	7.40e-12	1.88	1.31	5.10e-12	2.76e-02	0	4.94e+03
6	cage13	8	1.00	5.83e-12	2.30	1.39	5.04e-12	4.01e-02	1.87e-08	1.21e+05
7	cage14	8	1.00	9.20e-12	2.33	1.40	5.12e-12	1.51e-02	1.80e-07	2.53e+04
8	cage15	8	1.00	8.18e-12	2.35	1.40	5.08e-12	1.92e-02	3.20e-07	2.13e+05
9	crashbasis	10	1.00	3.83e-11	1.92	1.31	5.01e-12	2.73e-03	0	1.27e+03
10	dc1	11	1.00	2.13e-11	1.79	1.28	2.88e-12	4.08e-07	1.53e-06	2.10e+00
11	dc2	9	1.00	2.13e-11	1.62	1.24	3.19e-12	5.14e-06	0	2.28e-01
12	dc3	31	1.00	5.93e-12	1.72	1.26	3.18e-12	2.37e-04	0	2.45e+04
13	Goodwin_095	120	1.00	9.95e-11	1.81	1.29	4.50e-12	1.47e-03	0	1.15e+04
14	Goodwin_127	159	1.00	9.85e-11	1.72	1.26	4.52e-12	1.59e-03	0	1.37e+04
15	hcircuit	31	1.03	5.23e-11	1.81	1.26	7.16e-12	2.07e-04	0	1.10e+04
16	language	9	1.00	3.34e-11	1.87	1.30	5.52e-13	3.34e-04	8.89e-10	4.07e+02
17	majorbasis	10	1.00	2.43e-11	1.93	1.32	7.18e-12	4.73e-03	0	5.49e+02
18	memchip	9	1.00	4.69e-11	1.98	1.33	6.07e-12	1.54e-05	9.77e-07	1.84e+04
19	ML_Laplace	20	1.00	4.40e-11	1.82	1.29	3.77e-12	1.90e-05	2.92e-09	3.95e+03
20	rajat31	17	1.00	6.10e-11	2.42	1.41	7.20e-12	5.68e-05	7.92e-05	1.85e+06
21	ss	28	1.00	9.20e-11	2.03	1.34	2.21e-12	1.12e-04	9.42e-08	2.18e+05
22	ss1	7	1.00	2.77e-11	1.82	1.29	4.43e-12	9.68e-04	0	4.54e+03
23	stomach	10	1.00	2.75e-11	2.16	1.37	4.45e-12	4.93e-03	0	3.89e+04
24	torso2	9	1.00	8.89e-11	1.81	1.29	5.66e-12	1.81e-03	0	6.82e+02
25	trans5	11	1.00	1.20e-11	2.00	1.33	4.88e-12	4.22e-05	1.46e-06	1.21e+04
26	Transport	28	1.00	9.29e-11	2.86	1.48	6.26e-12	1.94e-05	7.02e-09	1.30e+03
27	vas_stokes_1M	72	1.00	8.43e-11	2.26	1.39	1.87e-12	1.26e-04	2.63e-07	1.21e+04
28	vas_stokes_2M	65	1.00	8.70e-11	2.29	1.39	1.86e-12	1.15e-04	4.02e-07	1.39e+04
29	xenon2	22	1.00	8.94e-11	1.59	1.23	5.33e-12	1.99e-05	0	1.53e+02

Table 5: Results from cFGMRES-GMRES with the relaxed compression strategy. While we get slightly better compression than when using the base compression strategy, the value of  $\zeta_k$  only grows in the final iterations, and thus the total compression ratio remains small overall.

	name	iter	rel iter	$\eta_b$	$\rho$	$\mu$	min $\zeta_k$	max $\zeta_k$	min $\tilde{\varphi}_k$	max $\tilde{\varphi}_k$
1	atmosmodd	23	2.09	1.75e-06	13.93	0.92	2.34e-03	6.11e-01	3.83e+02	6.70e+05
2	atmosmodj	23	2.09	1.75e-06	13.02	0.92	2.45e-03	6.11e-01	1.26e+03	8.83e+05
3	atmosmodl	21	2.10	3.12e-06	14.86	0.92	1.52e-02	5.95e-01	2.11e+03	1.12e+07
4	atmosmodm	21	2.10	3.19e-05	13.86	0.92	2.66e-02	7.33e-01	1.14e+03	5.19e+06
5	cage12	17	2.12	2.09e-06	15.76	0.91	5.10e-01	5.15e-01	1.04e+02	6.48e+04
6	cage13	17	2.12	1.55e-06	15.61	0.91	5.04e-01	5.10e-01	2.06e+03	1.49e+06
7	cage14	17	2.12	2.03e-06	15.87	0.91	5.12e-01	5.18e-01	3.89e+03	6.55e+06
8	cage15	17	2.12	1.86e-06	15.75	0.91	5.08e-01	5.14e-01	3.16e+03	1.12e+07
9	crashbasis	21	2.10	4.41e-08	15.52	0.92	3.94e-01	5.01e-01	4.44e+03	4.04e+05
10	dc1	17	1.55	4.33e-12	2.64	1.04	2.82e-05	2.88e-01	1.10e+00	2.87e+05
11	dc2	18	2.00	5.72e-12	4.17	0.89	9.34e-04	3.19e-01	2.35e+03	4.81e+06
12	dc3	36	1.16	9.48e-12	6.90	1.53	1.53e-03	3.18e-01	1.21e+03	1.04e+06
13	Goodwin_095	138	1.15	9.51e-11	11.53	1.62	1.69e-02	4.50e-01	7.29e+00	8.45e+05
14	Goodwin_127	183	1.15	9.61e-11	11.56	1.62	1.73e-02	4.52e-01	2.38e+02	4.59e+06
15	hcircuit	57	1.90	4.19e-11	13.27	1.01	2.51e-03	7.16e-01	5.61e+03	1.87e+07
16	language	11	1.22	1.08e-11	9.85	1.51	1.09e-02	5.52e-02	5.68e+03	1.83e+05
17	majorbasis	21	2.10	5.95e-06	17.87	0.93	4.74e-01	7.18e-01	2.93e+03	1.77e+06
18	memchip	19	2.11	1.07e-06	8.32	0.90	4.69e-03	6.07e-01	9.86e+04	9.94e+06
19	ML_Laplace	31	1.55	7.47e-11	4.05	1.11	2.83e-04	3.77e-01	1.04e+02	6.88e+05
20	rajat31	35	2.06	9.36e-10	9.33	0.92	1.44e-03	7.20e-01	1.12e+04	4.71e+07
21	ss	37	1.32	6.51e-11	6.34	1.35	1.45e-03	2.21e-01	8.51e+03	2.69e+07
22	ss1	15	2.14	2.21e-06	14.68	0.90	4.36e-01	4.43e-01	6.54e+01	1.62e+06
23	stomach	21	2.10	3.66e-08	14.77	0.92	4.31e-01	4.48e-01	1.22e+03	8.61e+05
24	torso2	19	2.11	3.32e-06	16.13	0.92	5.33e-01	5.66e-01	1.37e+03	2.56e+06
25	trans5	21	1.91	1.39e-11	6.65	0.97	4.33e-04	4.88e-01	2.06e+03	4.38e+05
26	Transport	57	2.04	1.05e-09	13.10	0.95	4.83e-04	6.26e-01	1.46e+02	1.29e+06
27	vas_stokes_1M	77	1.07	6.60e-11	7.47	1.66	1.09e-03	1.87e-01	2.14e+03	1.16e+06
28	vas_stokes_2M	68	1.05	5.38e-11	7.32	1.69	1.52e-03	1.86e-01	3.82e+03	4.59e+06
29	xenon2	39	1.77	7.64e-11	3.37	0.97	3.95e-04	5.33e-01	2.98e+00	1.86e+05

Table 6: Results from cFGMRES-GMRES with the double relaxed compression strategy. This strategy appears to allow for compression errors that are too large, which results in extra iterations for most of the matrices, negating the positive effect of the compressing  $Z$  ( $\rho$ ) on the overall memory used by cFGMRES ( $\mu$ ).

	name	iter	rel iter	$\eta_b$	$\rho$	$\mu$	min $\zeta_k$	max $\zeta_k$	min $\tilde{\varphi}_k$	max $\tilde{\varphi}_k$
1	atmosmodd	12	1.09	2.13e-11	8.05	1.65	1.19e-04	5.83e-02	1.82e+02	4.62e+04
2	atmosmodj	12	1.09	4.97e-11	7.92	1.64	9.62e-05	5.82e-02	2.69e+01	2.64e+04
3	atmosmodl	10	1.00	9.14e-11	8.76	1.80	1.79e-04	5.77e-02	8.56e+01	4.88e+05
4	atmosmodm	11	1.10	1.22e-11	6.87	1.61	4.03e-04	6.87e-02	1.18e+02	2.39e+04
5	cage12	8	1.00	9.37e-11	10.49	1.83	1.45e-02	5.08e-02	2.38e+02	6.67e+05
6	cage13	8	1.00	6.33e-11	10.53	1.83	1.29e-02	5.13e-02	8.47e+03	4.16e+04
7	cage14	8	1.00	4.63e-11	10.52	1.83	1.17e-02	4.36e-02	8.78e+03	7.16e+05
8	cage15	8	1.00	3.20e-11	10.46	1.83	1.06e-02	4.16e-02	2.59e+04	1.09e+06
9	crashbasis	10	1.00	9.75e-11	12.16	1.85	1.00e-02	4.67e-02	6.06e+02	4.42e+04
10	dc1	11	1.00	2.02e-11	1.39	1.16	3.01e-06	2.53e-02	0	5.01e+04
11	dc2	9	1.00	2.98e-11	3.40	1.55	4.23e-05	2.95e-02	2.27e+00	1.55e+06
12	dc3	31	1.00	1.40e-11	5.92	1.71	7.96e-04	1.76e-02	5.24e+02	2.60e+05
13	Goodwin_095	118	0.98	9.69e-11	10.31	1.85	6.40e-03	4.13e-02	5.99e+01	8.49e+06
14	Goodwin_127	158	0.99	8.99e-11	10.49	1.84	6.29e-03	4.18e-02	8.30e+01	1.45e+06
15	hcircuit	36	1.20	5.45e-11	10.48	1.54	9.68e-04	7.15e-02	1.31e+03	8.06e+05
16	language	9	1.00	4.21e-11	7.27	1.76	8.17e-04	5.02e-03	7.61e+02	1.06e+05
17	majorbasis	11	1.10	1.85e-11	11.01	1.68	1.03e-02	5.30e-02	5.51e+02	3.61e+04
18	memchip	10	1.11	2.46e-11	6.11	1.57	1.00e-04	3.72e-02	6.80e+03	2.94e+05
19	ML_Laplace	21	1.05	4.18e-11	3.96	1.54	9.40e-05	2.53e-02	2.83e+01	6.59e+04
20	rajat31	18	1.06	8.01e-11	8.31	1.70	1.89e-04	6.61e-02	8.38e+04	1.55e+06
21	ss	30	1.07	6.11e-11	6.49	1.63	6.82e-04	1.88e-02	5.01e+03	1.51e+06
22	ss1	7	1.00	1.23e-11	8.92	1.80	3.30e-03	3.88e-02	4.08e+02	2.26e+04
23	stomach	10	1.00	5.56e-11	12.94	1.86	1.70e-02	4.11e-02	1.53e+03	7.18e+04
24	torso2	10	1.11	1.14e-11	10.66	1.66	1.52e-02	4.64e-02	1.02e+03	2.67e+05
25	trans5	11	1.00	1.93e-11	5.93	1.71	7.24e-05	4.65e-02	5.40e+02	1.07e+05
26	Transport	32	1.14	7.76e-11	8.43	1.59	1.13e-04	6.00e-02	6.77e+01	1.08e+05
27	vas_stokes_1M	73	1.01	8.37e-11	6.84	1.72	6.76e-04	1.38e-02	8.56e+02	6.50e+05
28	vas_stokes_2M	65	1.00	8.97e-11	6.65	1.74	7.11e-04	1.74e-02	5.00e+03	9.54e+06
29	xenon2	24	1.09	7.11e-11	3.32	1.44	1.10e-04	3.60e-02	9.16e+00	7.74e+04

Table 7: Results from cFGMRES-GMRES with the equal compression strategy. This strategy finds a good balance between the size of the compression error ( $\zeta_k$ ) and the resulting compression ratios on  $Z$  ( $\rho$ ).

## A Exploiting dimension information

In our experiments  $z_k$  is an approximate solution of  $Az = v_k$  and we only looked at  $z_k$  as a vector. However, if the linear system  $Ax = b$  models the solution of some higher dimensional problem, e.g., a 2D or 3D problem, then we can look at  $z_k$  in this multidimensional space. The SZ compressor can exploit this information by modifying its prediction schemes to take into account the fact that the data is multidimensional and – possibly – get higher compression ratios.

To study this effect, we look at what happens to the compression and memory ratios when we pass the physical dimensions of the linear system to the SZ compressor. Consider therefore the test problem from [24]:

$$-\Delta u + \gamma(xu_x + yu_y) + \beta = f$$

on  $[0, 1]^2$  with Dirichlet boundary conditions. If we discretize this problem using finite differences on a regular grid with 2048 internal points in each dimension, then  $x$  and the  $z_k$  are vectors of length  $2048^2$ , but we can interpret them on a  $2048 \times 2048$  grid. We solve this problem using cFGMRES-GMRES with the equal compression strategy and the same parameters as described in section 6, but we do not limit the number of iterations for the preconditioner. Again, we take  $x^*$  a vector with random uniform entries drawn from  $[-1, 1]$ , and consider the following parameters:

$$\beta \in \{-100, -10, 10, 100\} \quad \text{and} \quad \gamma \in \{10, 100, 1000\}.$$

In Table 8 we report the results using both the original system and the scaled system. The scaling itself, however, seems to have little effect on the convergence for these matrices. The effect of passing the dimensions to SZ also seems to depend significantly on the parameter  $\gamma$ . For small values, i.e., diffusion dominated problems, the compression ratio increases slightly, but the effect on the memory ratio is negligible in all cases. This is consistent with observations of other applications of SZ: sometimes the dimension information is useful, sometimes it is not.

$\beta$	$\gamma$	dim info	original			scaled		
			iter	$\rho$	$\mu$	iter	$\rho$	$\mu$
-100	10	N	11	8.07	1.63	11	8.09	1.63
-100	10	Y	11	11.08	1.68	11	11.05	1.68
-100	100	N	11	9.23	1.66	11	9.20	1.65
-100	100	Y	11	11.18	1.68	11	11.20	1.68
-100	1000	N	11	11.07	1.68	11	11.10	1.68
-100	1000	Y	11	11.31	1.68	11	11.35	1.68
-10	10	N	11	8.04	1.63	11	8.08	1.63
-10	10	Y	11	11.11	1.68	11	11.11	1.68
-10	100	N	11	9.23	1.66	11	9.25	1.66
-10	100	Y	11	11.21	1.68	11	11.18	1.68
-10	1000	N	11	11.06	1.68	11	11.13	1.68
-10	1000	Y	11	11.34	1.68	11	11.40	1.68
10	10	N	11	8.12	1.64	11	8.08	1.63
10	10	Y	11	11.11	1.68	11	11.11	1.68
10	100	N	11	9.25	1.66	11	9.23	1.66
10	100	Y	11	11.17	1.68	11	11.19	1.68
10	1000	N	11	11.10	1.68	11	11.09	1.68
10	1000	Y	11	11.35	1.68	11	11.35	1.68
100	10	N	11	8.05	1.63	11	8.05	1.63
100	10	Y	11	11.14	1.68	11	11.11	1.68
100	100	N	11	9.29	1.66	11	9.26	1.66
100	100	Y	11	11.18	1.68	11	11.21	1.68
100	1000	N	11	11.06	1.68	11	11.11	1.68
100	1000	Y	11	11.31	1.68	11	11.36	1.68

Table 8: The dimension information only has a small impact on the compression ratios  $\rho$  for  $Z$ , most clearly visible for small values of  $\gamma$ . The effect on the memory ratio  $\mu$ , however, is almost negligible in all cases.



## B Pointwise vs. normwise

For the experiments in [section 6](#) we always used the normwise error control from SZ. In order to control the maximum normwise relative error, we passed the value  $\chi_k = \zeta_k \|z_k\|$  to SZ. We can also directly pass the value  $\chi_k = \zeta_k$  to SZ as a maximum pointwise relative error, which should result in stricter error control and hence lower compression ratios. This can be seen in [Table 9](#). There are some exceptions to this behaviour, which seem to occur with the matrices where the compression strategy has trouble compressing the data no matter which one of the two approaches is used. Similar results can also be observed for the other compression strategies, see [Appendix D](#).

	name	pointwise error control				normwise error control			
		iter	$\eta_b$	$\rho$	$\mu$	iter	$\eta_b$	$\rho$	$\mu$
1	atmosmodd	11	4.48e-11	5.98	1.71	12	2.13e-11	8.05	1.65
2	atmosmodj	12	1.09e-11	5.47	1.57	12	4.97e-11	7.92	1.64
3	atmosmodl	10	2.60e-11	6.01	1.71	10	9.14e-11	8.76	1.80
4	atmosmodm	10	3.02e-11	5.53	1.69	11	1.22e-11	6.87	1.61
5	cage12	8	1.63e-11	6.95	1.75	8	9.37e-11	10.49	1.83
6	cage13	8	1.20e-11	7.02	1.75	8	6.33e-11	10.53	1.83
7	cage14	8	2.87e-11	7.21	1.76	8	4.63e-11	10.52	1.83
8	cage15	8	2.00e-11	7.15	1.75	8	3.20e-11	10.46	1.83
9	crashbasis	10	5.71e-11	7.39	1.76	10	9.75e-11	12.16	1.85
10	dc1	11	1.84e-11	2.71	1.46	11	2.02e-11	1.39	1.16
11	dc2	9	1.23e-11	2.09	1.35	9	2.98e-11	3.40	1.55
12	dc3	31	1.02e-11	3.87	1.59	31	1.40e-11	5.92	1.71
13	Goodwin_095	119	9.29e-11	6.58	1.75	118	9.69e-11	10.31	1.85
14	Goodwin_127	157	9.57e-11	6.95	1.77	158	8.99e-11	10.49	1.84
15	hcircuit	32	6.43e-11	3.66	1.49	36	5.45e-11	10.48	1.54
16	language	9	2.98e-11	4.71	1.65	9	4.21e-11	7.27	1.76
17	majorbasis	10	8.04e-11	7.25	1.76	11	1.85e-11	11.01	1.68
18	memchip	10	7.70e-12	3.67	1.45	10	2.46e-11	6.11	1.57
19	ML_Laplace	20	4.95e-11	3.10	1.51	21	4.18e-11	3.96	1.54
20	rajat31	17	6.14e-11	5.17	1.68	18	8.01e-11	8.31	1.70
21	ss	28	9.58e-11	5.16	1.68	30	6.11e-11	6.49	1.63
22	ss1	6	7.52e-11	6.50	1.98	7	1.23e-11	8.92	1.80
23	stomach	10	3.75e-11	7.11	1.75	10	5.56e-11	12.94	1.86
24	torso2	10	1.12e-11	6.57	1.58	10	1.14e-11	10.66	1.66
25	trans5	11	4.29e-12	1.64	1.24	11	1.93e-11	5.93	1.71
26	Transport	29	5.90e-11	6.36	1.68	32	7.76e-11	8.43	1.59
27	vas_stokes_1M	72	6.78e-11	5.19	1.68	73	8.37e-11	6.84	1.72
28	vas_stokes_2M	65	7.71e-11	5.31	1.68	65	8.97e-11	6.65	1.74
29	xenon2	22	8.29e-11	2.61	1.45	24	7.11e-11	3.32	1.44

Table 9: Pointwise relative vs. normwise relative compression using the equal compression strategy.

## C Detailed results for some of the matrices

Some detailed results for the `cake12`, `Transport`, and `xenon` matrices are shown in figures 3–5. We can see how the compression ratio  $\rho$  for  $Z$ , the normwise relative compression error  $\zeta$  (or  $\tilde{\zeta}$ ), and the maximum pointwise relative compression error  $\tilde{\varphi}$  vary in each iteration.

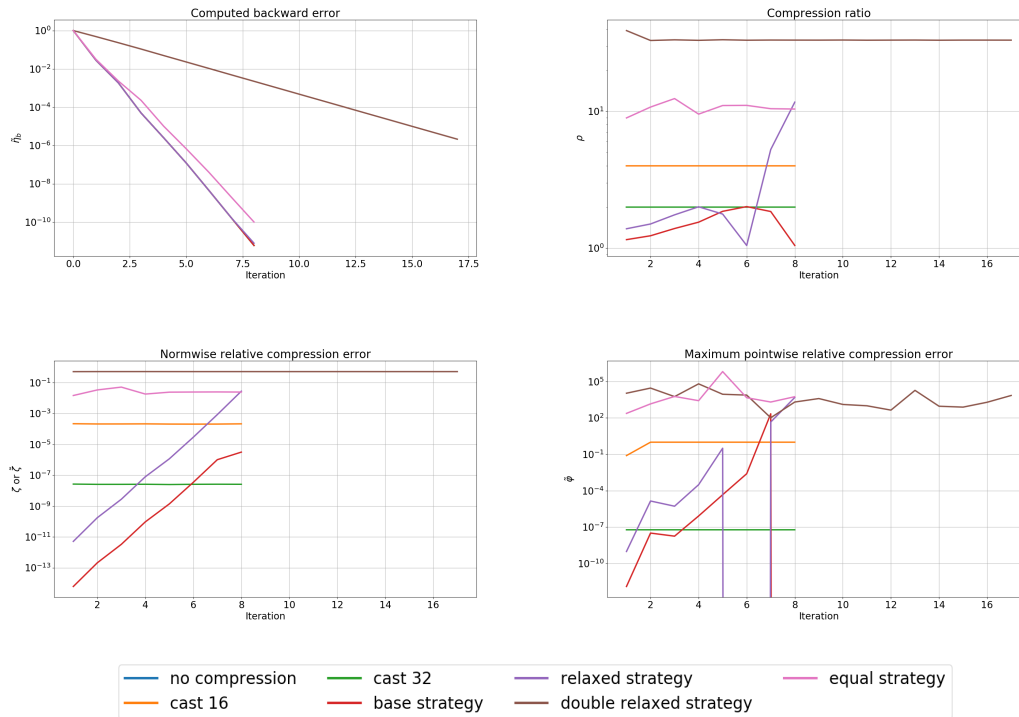


Figure 3: Detailed results for the `cake12` matrix.

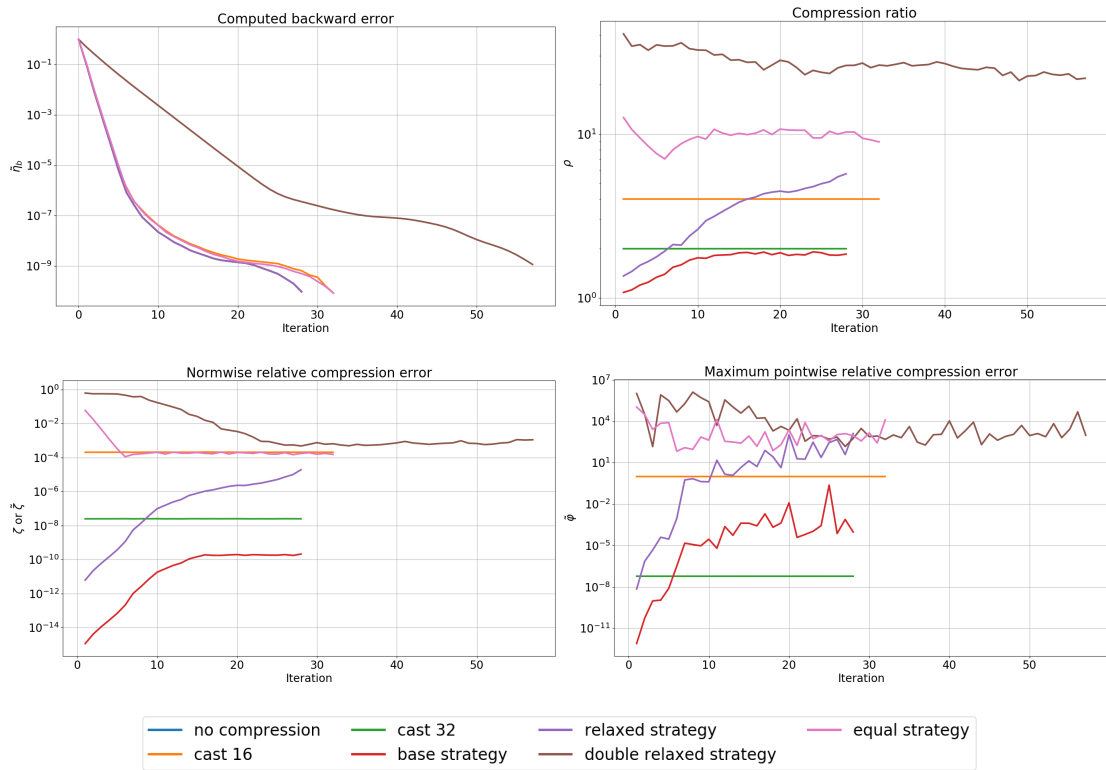


Figure 4: Detailed results for the Transport matrix.

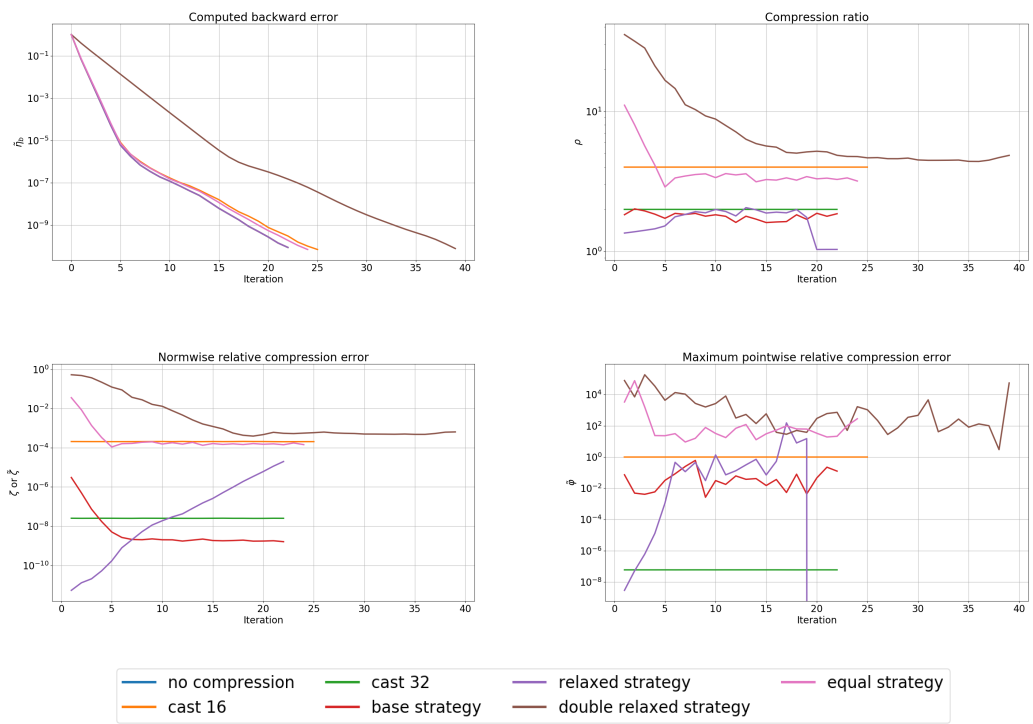


Figure 5: Detailed results for the xenon matrix.

## D Full tables

All results presented in the main text or in the previous appendices are summaries of the full results from our numerical experiments. What follows are the full unabridged results from the different experiments we performed. This includes the results for all the different compression strategies we considered for both the experiments with the original and the scaled version of the matrices.

Some remarks:

- [Tables 10–29](#) hold the results from the experiments on the original systems.
- [Tables 30–49](#) hold the results from the experiments on the scaled systems.
- For each compression strategy using SZ, we consider both the pointwise relative and normwise error control, passing  $\zeta_k$  and  $\zeta_k \|z_k\|$  respectively to SZ.
- We also consider a number of additional compression strategies, which are described in the next part of this section.
- Figures analogous to [figures 1](#) and [2](#) for all the compression strategies are given by [figures 6](#) and [7](#) for the applications to the original systems, and by [figures 8](#) and [9](#) for the applications to the scaled systems.

### D.1 Backtracking strategy

If FGMRES without compression converges with residuals

$$\|p_k\| = \|v_k - Az_k\|,$$

then we would expect to see similar convergence with other vectors  $\tilde{z}_k$  for which

$$\|v_k - Az_k\| \approx \|v_k - A\tilde{z}_k\|.$$

We can therefore loop over the values  $\zeta_k \in \{10^{-1}, 10^{-2}, \dots, 10^{-18}\}$ , determine  $\tilde{z}_k$  by compressing and decompressing  $z_k$  and find the largest  $\zeta_k$  such that

$$\|v_k - A\tilde{z}_k\| \leq (1 + \tau) \|v_k - Az_k\|.$$

Here,  $0 < \tau$  is a small tolerance value, which we take equal to 0.05 in our numerical experiments.

While we see in our experimental results that this strategy can find good values for the compression ratio  $\rho$  of  $Z$  and the memory ratio  $\mu$ , it is a computationally expensive strategy. This is because for each value of  $\zeta_k$  we need to try, we require an additional matrix vector product with  $A$  in order to calculate the new residual. Looking at the values of  $\zeta_k$  in the tables, this can sometimes mean an additional 2 to 5 matrix vector products per iteration.

### D.2 Heuristic strategy

The principle conclusion in the theory of inexact Krylov subspace methods is that the inexactness in the matrix vector product can grow as the iterations proceed. Translated to our context, this means that the compression error in each iteration can grow. This behaviour can for example be observed in the base and relaxed compression strategies. Because these two strategies do not seem to be very effective at compressing the search space  $Z$ , we could also try to systematically increase the allowed compression error in each iteration. More precisely, if FGMRES without

compression converges in  $\ell_{ref}$  iterations, then we will start with a value of  $\zeta_k = 1e-8$  and increase it by one order of magnitude every 10% of  $\ell_{ref}$ .

While this strategy seems to give good results for some matrices, it performs poorly on others. It is likely that the increase of  $\zeta_k$  after 10% of  $\ell_{ref}$  is too fast for some matrices. Optimizing the speed at which to increase  $\zeta_k$  for each matrix separately could make this strategy more viable. Depending on the application this, however, may or may not be a possibility.

### D.3 SZ 16 & SZ 32

By generating vectors  $z \in \mathbb{R}^n$  for different values of  $n$  and casting them to 16 or 32 bit, we can empirically approximate the maximum normwise relative error and maximum pointwise relative error for the casting strategies. We will use the following approximations

- **cast 16:**  $\tilde{\zeta} = 3.26e-4$  and  $\tilde{\varphi} = 1$
- **cast 32:**  $\tilde{\zeta} = 4.17e-8$  and  $\tilde{\varphi} = 5.96e-8$ .

Depending on whether we're using the normwise relative or pointwise relative mode we can therefore pass either  $\chi_k = \zeta \|z_k\|$  or  $\chi_k = \varphi$  to SZ in order to mimic the cast 16 and cast 32 compression strategies with SZ.

We added the normwise relative mode for completeness, but this one cannot be compared to its casting counterparts, since those errors are pointwise. For the pointwise mode, we see that the SZ 32 strategy performs similarly to its casting counterpart. For the SZ 16 strategy this is, however, not the case. While the compression ratios  $\rho$  for  $Z$  are bigger than 4 most of the time, the algorithm does not converge for most matrices, resulting in  $\mu < 1$ . A possible explanation for this is the value  $\tilde{\varphi} = 1$  that we used. While it is true that casting a 64 bit vector to a 16 bit vector can result in all information being lost in certain components of the vector, it is unlikely that this will happen in all components. Passing a maximum pointwise relative error  $\tilde{\varphi} = 1$  to SZ, however, will allow for this to occur in every component.





















	name	iter	rel iter	$\eta_b$	$\rho$	$\mu$	$\min \chi_k$	$\max \chi_k$	$\min \zeta_k$	$\max \zeta_k$	$\min \tilde{\zeta}_k$	$\max \tilde{\zeta}_k$	$\min \tilde{\varphi}_k$	$\max \tilde{\varphi}_k$
1	atmosmodd	11	1.0	1.32e-11	5.39	1.69	4.17e-08	4.17e-08	3.91e-06	0.00488	3.91e-06	0.00488	11.0	1.19e+05
2	atmosmodj	11	1.0	8.83e-11	5.3	1.68	4.17e-08	4.17e-08	3.64e-06	0.00488	3.64e-06	0.00488	1.57	5.93e+03
3	atmosmodl	10	1.0	1.61e-11	6.75	1.74	4.17e-08	4.17e-08	3.57e-05	0.0154	3.57e-05	0.0154	1.11e+02	1.03e+04
4	atmosmodm	14	1.4	5.79e-11	7.27	1.3	4.17e-08	4.17e-08	0.00184	0.195	0.00184	0.195	1.65e+03	9.03e+04
5	cage12	8	1.0	3.46e-11	1.97	1.33	4.17e-08	4.17e-08	7.07e-09	2.9e-08	9.17e-10	1.21e-08	7.19e-06	0.00353
6	cage13	8	1.0	5.38e-11	1.94	1.32	4.17e-08	4.17e-08	6.35e-09	2.91e-08	6.3e-10	5.93e-09	1.77e-05	0.0051
7	cage14	8	1.0	5.28e-11	1.93	1.32	4.17e-08	4.17e-08	6.01e-09	2.92e-08	6.79e-10	1.21e-08	7.7e-05	0.0242
8	cage15	8	1.0	9.45e-11	1.95	1.32	4.17e-08	4.17e-08	5.45e-09	2.92e-08	5.01e-10	1.29e-08	0.000275	0.171
9	crashbasis	10	1.0	3.15e-11	1.04	1.02	4.17e-08	4.17e-08	1.14e-07	1.35e-05	0.0	0.0	0.0	0.0
10	dc1	155	1.12	9.96e-11	1.91	1.22	4.17e-08	4.17e-08	1.03e-09	4.05e-07	0.0	1.54e-07	0.0	3.13
11	dc2	139	1.56	8.21e-11	1.4	0.879	4.17e-08	4.17e-08	1.9e-09	3.93e-07	0.0	1.04e-07	0.0	50.1
12	dc3	145	1.11	7.36e-11	1.97	1.24	4.17e-08	4.17e-08	2.02e-09	5.04e-07	0.0	2.03e-07	0.0	34.8
13	Goodwin_095	244	0.996	8.92e-11	1.65	1.25	4.17e-08	4.17e-08	1.72e-11	1.87e-08	1.49e-12	3.18e-09	7.65e-07	0.0177
14	Goodwin_127	169	1.0	9.8e-11	1.66	1.25	4.17e-08	4.17e-08	1.38e-11	1.85e-08	1.14e-12	3.28e-09	2.1e-06	0.0656
15	hcircuit	220	1.02	9.87e-11	1.73	1.25	4.17e-08	4.17e-08	4.35e-09	1.48e-06	0.0	7.44e-09	0.0	0.0152
16	language	9	1.0	3.4e-11	2.03	1.34	4.17e-08	4.17e-08	1.02e-08	5.87e-08	7.66e-10	6.83e-09	0.000269	0.394
17	majorbasis	10	1.0	4.67e-11	1.89	1.31	4.17e-08	4.17e-08	6.6e-08	6.48e-07	1.14e-08	1.7e-07	0.000525	0.113
18	memchip	68	1.0	9.05e-11	1.72	1.26	4.17e-08	4.17e-08	7.75e-10	1.21e-07	1.82e-11	2.9e-08	1.37e-05	0.611
19	ML_Laplace	66	1.25	7.84e-11	4.28	1.35	4.17e-08	4.17e-08	0.000298	0.371	0.000298	0.365	23.6	7.68e+09
20	rajat31	26	1.0	5.26e-11	2.03	1.34	4.17e-08	4.17e-08	1.12e-09	1.38e-06	2.32e-10	9.29e-07	3.84e-05	3.35e+03
21	ss	10	1.0	5.62e-11	1.7	1.26	4.17e-08	4.17e-08	3.07e-10	5.79e-08	1.4e-11	5.26e-09	0.000608	1.76
22	ss1	7	1.0	2.74e-11	1.89	1.31	4.17e-08	4.17e-08	1.41e-08	5.03e-08	1.91e-09	1.17e-08	0.000152	0.0406
23	stomach	10	1.0	4e-11	2.01	1.33	4.17e-08	4.17e-08	5.7e-09	3.94e-08	9.23e-10	6.72e-09	0.000224	0.0434
24	torso2	10	1.0	2.6e-11	2.03	1.34	4.17e-08	4.17e-08	1.96e-08	1.22e-07	3.13e-09	2.64e-08	0.00031	0.0133
25	trans5	440	1.06	9.85e-11	1.99	1.28	4.17e-08	4.17e-08	2.54e-09	5.23e-07	0.0	2.82e-07	0.0	6.11e+02
26	Transport	34	1.0	7.55e-11	1.62	1.24	4.17e-08	4.17e-08	7.22e-12	1.01e-08	7.28e-13	1.6e-09	1.88e-06	0.00712
27	vas_stokes_1M	76	1.0	7.47e-11	1.82	1.29	4.17e-08	4.17e-08	4.3e-10	6.59e-08	4.22e-11	1.46e-08	0.000385	4.67
28	vas_stokes_2M	72	1.0	5.94e-11	1.77	1.28	4.17e-08	4.17e-08	3.87e-10	6.1e-08	2.98e-11	7.29e-09	0.000285	12.4
29	xenon2	45	2.05	1.0	1.1e+04	0.978	4.17e-08	4.17e-08	4.67e+18	1.91e+21	0.727	1.99	1.59e+03	inf

Table 28: cFGMRES on the original system with the normwise SZ 32 compression strategy.



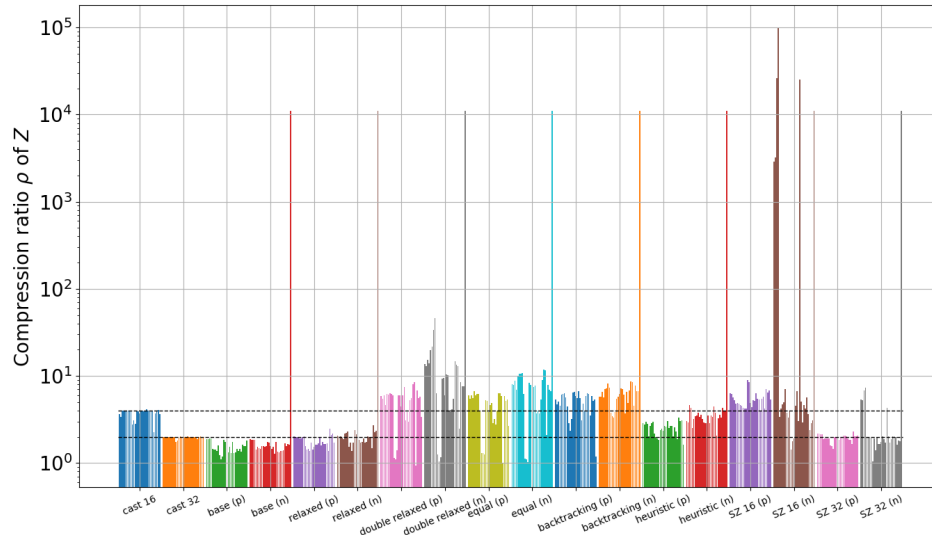


Figure 6: The compression ratio  $\rho$  of  $Z$  grouped per compression strategy for the solution of the original systems. Each bar per strategy corresponds to a different matrix according to its id in [Table 1](#). Horizontal reference lines are added at  $\rho = 2$  and  $4$ .

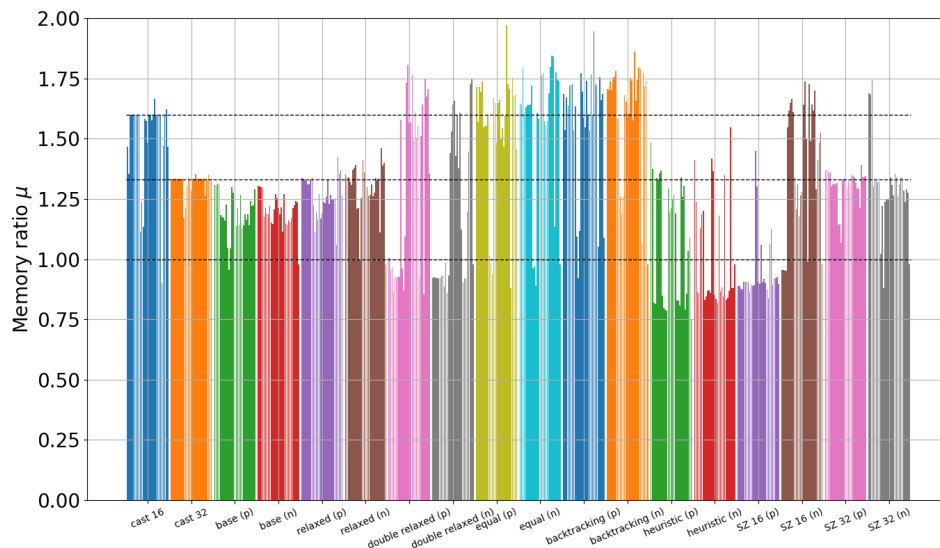


Figure 7: The memory ratio  $\mu$  grouped per compression strategy for the solution of the original systems. Each bar per strategy corresponds to a different matrix according to its id in [Table 1](#). Horizontal reference lines are added at  $\mu = 1$ ,  $1.33$ , and  $1.6$ .

$\beta$	$\gamma$	dim info	iter	$\rho$	$\mu$
-100	10	N	11	8.07	1.63
-100	10	Y	11	11.08	1.68
-100	50	N	11	9.01	1.65
-100	50	Y	11	11.07	1.68
-100	100	N	11	9.23	1.66
-100	100	Y	11	11.18	1.68
-100	500	N	10	11.46	1.84
-100	500	Y	10	12.28	1.85
-100	1000	N	11	11.07	1.68
-100	1000	Y	11	11.31	1.68
-50	10	N	11	8.09	1.63
-50	10	Y	11	11.11	1.68
-50	50	N	11	8.99	1.65
-50	50	Y	11	11.17	1.68
-50	100	N	11	9.24	1.66
-50	100	Y	11	11.19	1.68
-50	500	N	10	11.46	1.84
-50	500	Y	10	12.37	1.85
-50	1000	N	11	11.06	1.68
-50	1000	Y	11	11.34	1.68
-10	10	N	11	8.04	1.63
-10	10	Y	11	11.11	1.68
-10	50	N	11	8.98	1.65
-10	50	Y	11	11.16	1.68
-10	100	N	11	9.23	1.66
-10	100	Y	11	11.21	1.68
-10	500	N	10	11.49	1.84
-10	500	Y	10	12.37	1.85
-10	1000	N	11	11.06	1.68
-10	1000	Y	11	11.34	1.68
10	10	N	11	8.12	1.64
10	10	Y	11	11.11	1.68
10	50	N	11	8.96	1.65
10	50	Y	11	11.06	1.68
10	100	N	11	9.25	1.66
10	100	Y	11	11.17	1.68
10	500	N	10	11.44	1.84
10	500	Y	10	12.33	1.85
10	1000	N	11	11.10	1.68
10	1000	Y	11	11.35	1.68
50	10	N	11	8.03	1.63
50	10	Y	11	11.13	1.68
50	50	N	11	8.92	1.65
50	50	Y	11	11.17	1.68
50	100	N	11	9.32	1.66
50	100	Y	11	11.19	1.68
50	500	N	10	11.49	1.84
50	500	Y	10	12.28	1.85
50	1000	N	11	11.03	1.68
50	1000	Y	11	11.38	1.68
100	10	N	11	8.05	1.63
100	10	Y	11	11.14	1.68
100	50	N	11	8.93	1.65
100	50	Y	11	11.21	1.68
100	100	N	11	9.29	1.66
100	100	Y	11	11.18	1.68
100	500	N	10	11.48	1.84
100	500	Y	10	12.30	1.85
100	1000	N	11	11.06	1.68
100	1000	Y	11	11.31	1.68

Table 29: Results from the scaling experiment on the original system.















	name	iter	rel iter	$\eta_b$	$\rho$	$\mu$	$\min \chi_k$	$\max \chi_k$	$\min \zeta_k$	$\max \zeta_k$	$\min \tilde{\zeta}_k$	$\max \tilde{\zeta}_k$	$\min \tilde{\varphi}_k$	$\max \tilde{\varphi}_k$
1	atmosmodd	11	1.0	1.34e-11	5.78	1.7	0.00227	0.0119	1e-05	0.01	9.96e-06	0.01	20.5	1.31e+04
2	atmosmodj	11	1.0	7.66e-11	5.74	1.7	0.00227	0.0119	1e-05	0.01	9.98e-06	0.01	7.31	1.28e+04
3	atmosmodl	10	1.0	1.92e-11	6.6	1.74	0.00316	0.0257	1e-05	0.01	1e-05	0.00999	5.89	3.1e+04
4	atmosmodm	10	1.0	1.2e-11	5.68	1.7	0.00317	0.0141	0.0001	0.01	9.98e-05	0.01	1.34e+02	1.57e+04
5	cage12	8	1.0	8.47e-12	7.56	1.77	0.000986	0.0171	0.001	0.01	0.000994	0.01	22.8	2e+03
6	cage13	8	1.0	5.55e-12	7.88	1.77	0.000972	0.0162	0.001	0.01	0.000998	0.01	4.85e+02	6.59e+04
7	cage14	8	1.0	1.37e-11	8.49	1.79	0.00099	0.0155	0.001	0.01	0.001	0.01	2.31e+03	9.14e+05
8	cage15	8	1.0	1.01e-11	8.52	1.79	0.000991	0.0153	0.001	0.01	0.001	0.01	1.64e+04	1.28e+05
9	crashbasis	10	1.0	5.44e-11	7.91	1.78	0.00304	0.0269	0.001	0.01	0.000998	0.01	88.4	1.16e+04
10	dc1	11	1.0	1.68e-11	1.37	1.16	0.0014	0.129	1e-05	0.01	0.0	0.01	0.0	1.64e+04
11	dc2	9	1.0	2.11e-11	1.65	1.24	0.00143	0.00805	1e-05	0.01	0.0	0.00999	0.0	4.59e+04
12	dc3	30	0.968	7.38e-12	5.35	1.73	0.00365	0.124	0.0001	0.01	9.96e-05	0.01	34.4	1.6e+05
13	Goodwin_095	120	1.0	9.86e-11	5.79	1.71	0.00597	0.0189	0.001	0.01	0.000989	0.01	19.8	1.76e+05
14	Goodwin_127	160	1.01	8.9e-11	6.13	1.71	0.00562	0.0187	0.001	0.01	0.000993	0.01	7.2	2.54e+05
15	hcircuit	31	1.03	4.33e-11	6.64	1.69	0.00644	0.157	0.0001	0.01	9.92e-05	0.00997	2.88e+02	3.76e+05
16	language	10	1.11	1.04e-11	7.2	1.6	0.00194	0.0182	0.001	0.01	0.000996	0.01	4.34e+02	1.79e+05
17	majorbasis	10	1.0	2.37e-11	7.13	1.75	0.00178	0.00551	0.001	0.01	0.000998	0.00999	62.7	3.54e+04
18	memchip	10	1.11	4.85e-12	4.92	1.52	0.00337	0.0233	1e-05	0.01	9.98e-06	0.01	3.67e+03	1.15e+05
19	ML_Laplace	20	1.0	5.17e-11	3.45	1.55	0.00298	0.121	1e-05	0.01	7.17e-06	0.00998	51.5	1.97e+04
20	rajat31	16	0.941	8.61e-11	7.8	1.87	0.00287	0.0854	0.0001	0.01	9.66e-05	0.00995	2.18e+04	1.57e+06
21	ss	29	1.04	6.71e-11	6.19	1.67	0.00251	0.125	0.001	0.01	0.000993	0.00999	7.13e+02	3.24e+06
22	ss1	7	1.0	3.84e-11	7.31	1.76	0.000988	0.019	0.001	0.01	0.000995	0.01	53.5	1.02e+06
23	stomach	10	1.0	3.57e-11	11.3	1.84	0.00803	0.022	0.01	0.01	0.000992	0.01	2.12e+03	5.15e+04
24	torso2	10	1.11	7.24e-12	8.24	1.62	0.0023	0.0248	0.001	0.01	0.000995	0.01	3.45e+02	2.51e+04
25	trans5	11	1.0	1.46e-11	4.54	1.64	0.00343	0.084	1e-05	0.01	1e-05	0.01	5.27e+02	1.29e+05
26	Transport	28	1.0	7.91e-11	6.68	1.74	0.00229	0.183	1e-05	0.01	9.9e-06	0.00998	9.03	1.1e+04
27	vas_stokes_1M	73	1.01	6.52e-11	6.63	1.72	0.00243	0.12	0.001	0.01	0.000995	0.01	1.45e+03	1.21e+06
28	vas_stokes_2M	66	1.02	5.55e-11	6.51	1.71	0.00629	0.081	0.001	0.01	0.000997	0.01	1.47e+03	2.99e+06
29	xenon2	22	1.0	9.13e-11	2.77	1.47	0.00243	0.116	1e-05	0.01	0.0	0.01	0.0	6.49e+02

Table 42: cFGMRES on the scaled system with the normwise backtracking compression strategy.

	name	iter	rel iter	$\eta_b$	$\rho$	$\mu$	$\min \chi_k$	$\max \chi_k$	$\min \zeta_k$	$\max \zeta_k$	$\min \tilde{\zeta}_k$	$\max \tilde{\zeta}_k$	$\min \tilde{\varphi}_k$	$\max \tilde{\varphi}_k$
1	atmosmodd	11	1.0	9.85e-11	2.86	1.48	1e-08	0.001	1e-08	0.001	2.89e-09	0.000542	9.98e-09	0.000999
2	atmosmodj	12	1.09	4.2e-11	2.75	1.37	1e-08	0.001	1e-08	0.001	2.96e-09	0.000542	9.99e-09	0.001
3	atmosmodl	21	2.1	4.44e-09	2.98	0.821	1e-08	1.0	1e-08	1e+12	2.98e-09	0.321	9.99e-09	0.5
4	atmosmodm	21	2.1	1.98e-09	2.82	0.815	1e-08	1.0	1e-08	1e+12	2.89e-09	0.357	9.96e-09	0.5
5	cage12	8	1.0	1.04e-11	2.02	1.34	1e-08	0.1	1e-08	0.1	0.0	0.0535	0.0	0.0997
6	cage13	8	1.0	6.73e-12	2.76	1.47	1e-08	0.1	1e-08	0.1	2.98e-09	0.0531	9.97e-09	0.0997
7	cage14	8	1.0	1.72e-11	2.99	1.5	1e-08	0.1	1e-08	0.1	2.99e-09	0.053	1e-08	0.0997
8	cage15	8	1.0	1.37e-11	3.09	1.51	1e-08	0.1	1e-08	0.1	2.99e-09	0.0524	1e-08	0.0997
9	crashbasis	13	1.3	3.46e-11	2.61	1.19	1e-08	1.0	1e-08	1e+04	0.0	0.319	0.0	0.5
10	dc1	15	1.36	6.25e-12	1.61	1.01	1e-08	0.1	1e-08	0.1	0.0	0.0507	0.0	0.1
11	dc2	19	2.11	3.47e-08	1.96	0.763	1e-08	1.0	1e-08	1e+10	0.0	0.323	0.0	0.5
12	dc3	56	1.81	9.61e-12	1.88	0.856	1e-08	1.0	1e-08	1e+05	0.0	0.388	0.0	0.5
13	Goodwin_095	169	1.41	9.28e-11	2.18	1.07	1e-08	1.0	1e-08	1e+06	0.0	0.331	0.0	0.5
14	Goodwin_127	222	1.4	9.47e-11	2.24	1.09	1e-08	1.0	1e-08	1e+05	0.0	0.328	0.0	0.5
15	hcircuit	61	2.03	3.13e-09	1.91	0.782	1e-08	1.0	1e-08	1e+12	0.0	0.349	0.0	0.5
16	language	10	1.11	9.46e-11	2.74	1.35	1e-08	1.0	1e-08	10.0	1.91e-09	0.327	9.96e-09	0.5
17	majorbasis	13	1.3	5.15e-11	2.59	1.19	1e-08	1.0	1e-08	1e+04	0.0	0.32	0.0	0.5
18	memchip	19	2.11	7.55e-09	2.31	0.786	1e-08	1.0	1e-08	1e+10	2.8e-09	0.331	9.98e-09	0.5
19	ML_Laplace	41	2.05	2.87e-09	2.72	0.827	1e-08	1.0	1e-08	1e+12	2.84e-09	0.353	9.9e-09	0.5
20	rajat31	35	2.06	8.68e-10	2.62	0.819	1e-08	1.0	1e-08	1e+09	1.49e-09	0.38	1e-08	0.5
21	ss	48	1.71	8.15e-11	2.85	0.968	1e-08	1.0	1e-08	1e+07	2.96e-09	0.325	9.97e-09	0.5
22	ss1	7	1.0	3.95e-11	1.8	1.29	1e-08	0.01	1e-08	0.01	0.0	0.0053	0.0	0.01
23	stomach	12	1.2	2.69e-11	2.54	1.25	1e-08	1.0	1e-08	1e+03	0.0	0.364	0.0	0.5
24	torso2	11	1.22	4.93e-11	2.36	1.21	1e-08	1.0	1e-08	1e+02	0.0	0.319	0.0	0.5
25	trans5	11	1.0	1.67e-11	1.68	1.25	1e-08	0.001	1e-08	0.001	0.0	0.000489	0.0	0.000998
26	Transport	57	2.04	1.6e-09	3.42	0.859	1e-08	1.0	1e-08	1e+10	2.94e-09	0.332	9.99e-09	0.5
27	vas_stokes_1M	117	1.62	9.16e-11	2.85	1.01	1e-08	1.0	1e-08	1e+06	1.53e-09	0.33	9.98e-09	0.5
28	vas_stokes_2M	106	1.63	9.98e-11	3.06	1.02	1e-08	1.0	1e-08	1e+07	1.58e-09	0.326	9.97e-09	0.5
29	xenon2	45	2.05	2.75e-10	1.62	0.751	1e-08	1.0	1e-08	1e+06	0.0	0.32	0.0	0.5

Table 43: cFGMRES on the scaled system with the pointwise heuristic compression strategy.

	name	iter	rel iter	$\eta_b$	$\rho$	$\mu$	$\min \chi_k$	$\max \chi_k$	$\min \zeta_k$	$\max \zeta_k$	$\min \tilde{\zeta}_k$	$\max \tilde{\zeta}_k$	$\min \bar{\varphi}_k$	$\max \bar{\varphi}_k$
1	atmosmodd	12	1.09	2.34e-11	3.05	1.41	8.18e-09	0.398	1e-08	0.001	1.65e-09	0.000999	0.00014	4.06e+03
2	atmosmodj	13	1.18	8.01e-11	3.0	1.32	8.18e-09	1.0	1e-08	0.01	1.65e-09	0.00208	0.000224	9.22e+04
3	atmosmodl	21	2.1	2.28e-07	4.4	0.859	8.41e-09	1.0	1e-08	1e+12	3.27e-09	0.994	1.14e-05	2.29e+07
4	atmosmodm	21	2.1	3.73e-07	4.07	0.853	6.84e-09	1.0	1e-08	1e+12	3.19e-09	0.924	1.39e-05	1.17e+06
5	cage12	8	1.0	2.01e-11	2.53	1.43	9.86e-09	0.11	1e-08	0.1	0.0	0.1	0.0	7.47e+03
6	cage13	8	1.0	1.38e-11	3.12	1.51	9.88e-09	0.114	1e-08	0.1	3.32e-09	0.1	4.09e-05	1.05e+05
7	cage14	8	1.0	3.63e-11	3.34	1.54	9.9e-09	0.126	1e-08	0.1	3.28e-09	0.0998	0.000628	3.2e+05
8	cage15	8	1.0	2.81e-11	3.48	1.55	9.91e-09	0.128	1e-08	0.1	6.4e-09	0.0999	0.000505	2.33e+06
9	crashbasis	21	2.1	2.93e-09	2.8	0.814	7.77e-09	1.0	1e-08	1e+12	0.0	1.83	0.0	2.39e+06
10	dc1	15	1.36	1.61e-11	1.7	1.02	6.93e-09	1.0	1e-08	0.1	0.0	0.0998	0.0	7.12e+04
11	dc2	19	2.11	1.06e-08	2.84	0.812	6.91e-09	1.0	1e-08	1e+10	0.0	1.23	0.0	2.06e+06
12	dc3	63	2.03	4.85e-11	2.46	0.82	7.07e-09	1.0	1e-08	1e+07	0.0	1.28	0.0	3.61e+07
13	Goodwin_095	241	2.01	6.69e-09	3.0	0.854	6.08e-09	1.0	1e-08	1e+12	0.0	0.895	0.0	5.34e+06
14	Goodwin_127	319	2.01	1.02e-08	3.05	0.857	6.05e-09	1.0	1e-08	1e+11	0.0	0.894	0.0	5.92e+06
15	hcircuit	61	2.03	2.84e-08	3.64	0.866	6.44e-09	1.0	1e-08	1e+12	0.0	1.61	0.0	9.14e+06
16	language	19	2.11	3.14e-09	3.46	0.833	7.12e-09	1.0	1e-08	1e+10	4.52e-10	1.1	1.46e-05	8.84e+06
17	majorbasis	21	2.1	8.7e-09	3.68	0.843	5.51e-09	1.0	1e-08	1e+12	8.8e-10	2.23	4.23e-05	2.33e+06
18	memchip	19	2.11	1.8e-07	3.81	0.843	6.58e-09	1.0	1e-08	1e+10	2.64e-09	1.53	0.0123	4.65e+08
19	ML_Laplace	41	2.05	7.62e-09	3.54	0.857	7.04e-09	1.0	1e-08	1e+12	1.62e-09	1.81	8.42e-06	4.1e+07
20	rajat31	35	2.06	5.09e-09	4.98	0.885	6.59e-09	1.0	1e-08	1e+09	2.82e-09	1.31	0.112	3.37e+09
21	ss	57	2.04	9.56e-09	3.72	0.868	6.91e-09	1.0	1e-08	1e+10	7.93e-10	0.897	0.000574	6.04e+07
22	ss1	7	1.0	5.37e-11	2.24	1.38	9.88e-09	0.0219	1e-08	0.01	0.0	0.00992	0.0	5.84e+03
23	stomach	21	2.1	6.72e-09	3.57	0.84	8.81e-09	1.0	1e-08	1e+12	0.0	0.989	0.0	2e+07
24	torso2	19	2.11	3.62e-09	2.79	0.81	8.96e-09	1.0	1e-08	1e+10	0.0	0.961	0.0	2.89e+06
25	trans5	11	1.0	4.18e-11	2.51	1.43	5.81e-09	0.133	1e-08	0.001	0.0	0.000998	0.0	6.13e+03
26	Transport	57	2.04	1.82e-09	5.02	0.895	7.99e-09	1.0	1e-08	1e+10	1.61e-09	0.97	0.000157	4.17e+07
27	vas_stokes_1M	145	2.01	1.12e-09	3.77	0.878	5.95e-09	1.0	1e-08	1e+10	1.44e-09	0.816	0.000843	2.45e+07
28	vas_stokes_2M	131	2.02	5.68e-09	4.03	0.883	6.43e-09	1.0	1e-08	1e+10	9.18e-10	0.887	0.0029	2.74e+07
29	xenon2	45	2.05	3.03e-10	2.1	0.793	5.94e-09	1.0	1e-08	1e+06	0.0	1.03	0.0	7.26e+05

Table 44: cFGMRES on the scaled system with the normwise heuristic compression strategy.

	name	iter	rel iter	$\eta_b$	$\rho$	$\mu$	$\min \chi_k$	$\max \chi_k$	$\min \zeta_k$	$\max \zeta_k$	$\min \tilde{\zeta}_k$	$\max \tilde{\zeta}_k$	$\min \bar{\varphi}_k$	$\max \bar{\varphi}_k$
1	atmosmodd	23	2.09	5.75e-06	6.41	0.89	1.0	1.0	1.0	1.0	0.293	0.329	0.5	0.5
2	atmosmodj	23	2.09	6.98e-06	6.33	0.889	1.0	1.0	1.0	1.0	0.291	0.334	0.5	0.5
3	atmosmodl	21	2.1	6.01e-06	6.18	0.884	1.0	1.0	1.0	1.0	0.299	0.332	0.5	0.5
4	atmosmodm	21	2.1	7.44e-06	5.49	0.876	1.0	1.0	1.0	1.0	0.278	0.351	0.5	0.5
5	cage12	15	1.88	7.48e-11	5.19	0.967	1.0	1.0	1.0	1.0	0.296	0.338	0.5	0.5
6	cage13	15	1.88	4.42e-11	5.22	0.968	1.0	1.0	1.0	1.0	0.283	0.326	0.5	0.5
7	cage14	15	1.88	4.26e-11	5.16	0.967	1.0	1.0	1.0	1.0	0.282	0.351	0.5	0.5
8	cage15	15	1.88	4.53e-11	4.98	0.964	1.0	1.0	1.0	1.0	0.294	0.355	0.5	0.5
9	crashbasis	19	1.9	7.23e-11	5.4	0.959	1.0	1.0	1.0	1.0	0.3	0.32	0.5	0.5
10	dc1	23	2.09	2.22e-05	4.65	0.867	1.0	1.0	1.0	1.0	0.297	0.331	0.5	0.5
11	dc2	19	2.11	8.68e-05	4.74	0.861	1.0	1.0	1.0	1.0	0.295	0.331	0.5	0.5
12	dc3	63	2.03	1.54e-07	5.25	0.9	1.0	1.0	1.0	1.0	0.267	0.361	0.5	0.5
13	Goodwin_095	241	2.01	1.45e-09	5.86	0.918	1.0	1.0	1.0	1.0	0.299	0.333	0.5	0.5
14	Goodwin_127	319	2.01	1.19e-09	6.29	0.924	1.0	1.0	1.0	1.0	0.295	0.33	0.5	0.5
15	hcircuit	61	2.03	1.73e-05	4.13	0.879	1.0	1.0	1.0	1.0	0.271	0.329	0.5	0.5
16	language	18	2.0	5.96e-11	4.3	0.896	1.0	1.0	1.0	1.0	0.278	0.336	0.5	0.5
17	majorbasis	21	2.1	6.4e-11	4.9	0.868	1.0	1.0	1.0	1.0	0.31	0.328	0.5	0.5
18	memchip	19	2.11	6.11e-05	4.47	0.857	1.0	1.0	1.0	1.0	0.205	0.325	0.5	0.5
19	ML_Laplace	41	2.05	1.19e-06	6.27	0.905	1.0	1.0	1.0	1.0	0.287	0.33	0.5	0.5
20	rajat31	35	2.06	1.46e-05	5.38	0.891	1.0	1.0	1.0	1.0	0.258	0.384	0.5	0.5
21	ss	57	2.04	3.21e-07	5.75	0.905	1.0	1.0	1.0	1.0	0.296	0.334	0.5	0.5
22	ss1	15	2.14	3.99e-10	4.01	0.836	1.0	1.0	1.0	1.0	0.289	0.364	0.5	0.5
23	stomach	17	1.7	4.08e-11	5.31	1.06	1.0	1.0	1.0	1.0	0.253	0.367	0.5	0.5
24	torso2	18	2.0	2.35e-11	4.54	0.901	1.0	1.0	1.0	1.0	0.288	0.347	0.5	0.5
25	trans5	23	2.09	2.78e-05	4.4	0.863	1.0	1.0	1.0	1.0	0.285	0.335	0.5	0.5
26	Transport	57	2.04	2.67e-07	6.74	0.916	1.0	1.0	1.0	1.0	0.288	0.353	0.5	0.5
27	vas_stokes_1M	145	2.01	4.31e-08	6.33	0.921	1.0	1.0	1.0	1.0	0.288	0.339	0.5	0.5
28	vas_stokes_2M	131	2.02	7.68e-08	6.73	0.924	1.0	1.0	1.0	1.0	0.287	0.335	0.5	0.5
29	xenon2	45	2.05	2.07e-06	5.25	0.894	1.0	1.0	1.0	1.0	0.302	0.324	0.5	0.5

Table 45: cFGMRES on the scaled system with the pointwise SZ 16 compression strategy.



	name	iter	rel iter	$\eta_b$	$\rho$	$\mu$	$\min \chi_k$	$\max \chi_k$	$\min \zeta_k$	$\max \zeta_k$	$\min \tilde{\zeta}_k$	$\max \tilde{\zeta}_k$	$\min \tilde{\varphi}_k$	$\max \tilde{\varphi}_k$
1	atmosmodd	11	1.0	1.38e-11	1.85	1.3	4.17e-08	4.17e-08	4.11e-11	5.1e-08	1.17e-11	2.62e-08	1.54e-05	0.00145
2	atmosmodj	11	1.0	8.43e-11	1.83	1.29	4.17e-08	4.17e-08	3.79e-11	5.1e-08	1.1e-11	2.61e-08	1.66e-05	0.00358
3	atmosmodl	10	1.0	1.37e-11	1.85	1.3	4.17e-08	4.17e-08	1.16e-10	4.96e-08	4.93e-11	2.62e-08	1.12e-06	0.00106
4	atmosmodm	10	1.0	1.15e-11	1.76	1.28	4.17e-08	4.17e-08	2.85e-10	6.1e-08	4.97e-11	2.56e-08	1.9e-06	0.00927
5	cage12	8	1.0	5.42e-12	1.95	1.32	4.17e-08	4.17e-08	2.38e-08	4.23e-08	2.4e-09	1.34e-08	1.17e-05	0.0043
6	cage13	8	1.0	3.07e-12	1.97	1.33	4.17e-08	4.17e-08	2.39e-08	4.29e-08	2.3e-09	1.34e-08	6.35e-05	0.0166
7	cage14	8	1.0	7.68e-12	2.02	1.34	4.17e-08	4.17e-08	2.48e-08	4.39e-08	3.53e-09	2.63e-08	0.000628	0.167
8	cage15	8	1.0	6.46e-12	1.95	1.32	4.17e-08	4.17e-08	2.56e-08	4.45e-08	1.97e-09	2.57e-08	0.00101	0.784
9	crashbasis	10	1.0	3.82e-11	1.89	1.31	4.17e-08	4.17e-08	1.19e-08	5.36e-08	1.12e-09	1.49e-08	2.96e-05	0.0101
10	dc1	11	1.0	2.12e-11	1.69	1.26	4.17e-08	4.17e-08	3.27e-12	6.01e-08	6.95e-14	3.33e-08	6.84e-06	0.421
11	dc2	9	1.0	2.13e-11	1.81	1.29	4.17e-08	4.17e-08	7e-11	6.03e-08	3.78e-12	3.33e-08	2.62e-06	0.767
12	dc3	31	1.0	2.43e-11	1.74	1.27	4.17e-08	4.17e-08	3.24e-10	5.9e-08	9.82e-12	1.79e-08	2.41e-07	2.65
13	Goodwin_095	120	1.0	9.93e-11	1.77	1.28	4.17e-08	4.17e-08	2.18e-09	6.86e-08	1.38e-10	1.38e-08	4.38e-06	0.676
14	Goodwin_127	159	1.0	9.83e-11	1.78	1.28	4.17e-08	4.17e-08	2.23e-09	6.89e-08	2.15e-10	1.41e-08	5.79e-06	0.455
15	hcircuit	31	1.03	6.01e-11	1.82	1.26	4.17e-08	4.17e-08	1.91e-10	6.47e-08	2.31e-11	8.58e-09	0.000363	0.12
16	language	9	1.0	3.34e-11	1.95	1.32	4.17e-08	4.17e-08	1.01e-08	5.86e-08	7.69e-10	9.74e-09	5.53e-05	0.191
17	majorbasis	10	1.0	2.36e-11	1.91	1.31	4.17e-08	4.17e-08	9.9e-09	7.57e-08	8.89e-10	1.01e-08	3.07e-05	0.00212
18	memchip	9	1.0	4.69e-11	1.89	1.31	4.17e-08	4.17e-08	7.52e-11	6.34e-08	2.51e-11	2.14e-08	0.00237	0.593
19	ML_Laplace	20	1.0	4.38e-11	1.58	1.22	4.17e-08	4.17e-08	3.4e-11	5.92e-08	1.19e-11	1.35e-08	0.000346	0.525
20	rajat31	17	1.0	6.19e-11	1.78	1.28	4.17e-08	4.17e-08	4.6e-11	6.33e-08	9.16e-12	5.25e-08	0.0863	23.8
21	ss	28	1.0	9.28e-11	1.72	1.26	4.17e-08	4.17e-08	3.26e-10	6.03e-08	2.64e-11	5.48e-09	0.00114	4.15e+02
22	ss1	7	1.0	2.73e-11	1.98	1.33	4.17e-08	4.17e-08	1.88e-08	5.12e-08	2.45e-09	1.18e-08	0.00122	0.0586
23	stomach	10	1.0	2.73e-11	2.04	1.34	4.17e-08	4.17e-08	1.87e-08	5.2e-08	3.03e-09	1.22e-08	0.000487	0.144
24	torso2	9	1.0	8.87e-11	2.01	1.33	4.17e-08	4.17e-08	1.47e-08	4.65e-08	1.54e-09	1.31e-08	5.75e-05	0.00617
25	trans5	11	1.0	1.2e-11	1.89	1.31	4.17e-08	4.17e-08	4.99e-11	7.18e-08	4.72e-12	2.64e-08	3.6e-05	0.16
26	Transport	28	1.0	9.25e-11	1.74	1.27	4.17e-08	4.17e-08	2.47e-11	5.22e-08	2.8e-12	7.09e-09	1.98e-06	0.0287
27	vas_stokes_1M	72	1.0	8.55e-11	1.81	1.29	4.17e-08	4.17e-08	3.41e-10	7.01e-08	3.4e-11	1.53e-08	0.00125	3.58
28	vas_stokes_2M	65	1.0	8.84e-11	1.77	1.28	4.17e-08	4.17e-08	4.94e-10	6.49e-08	4.63e-11	7.88e-09	0.00336	20.5
29	xenon2	22	1.0	8.94e-11	1.54	1.21	4.17e-08	4.17e-08	3.73e-11	7.01e-08	2.18e-12	1.35e-08	4.79e-06	0.091

Table 48: cFGMRES on the scaled system with the normwise SZ 32 compression strategy.

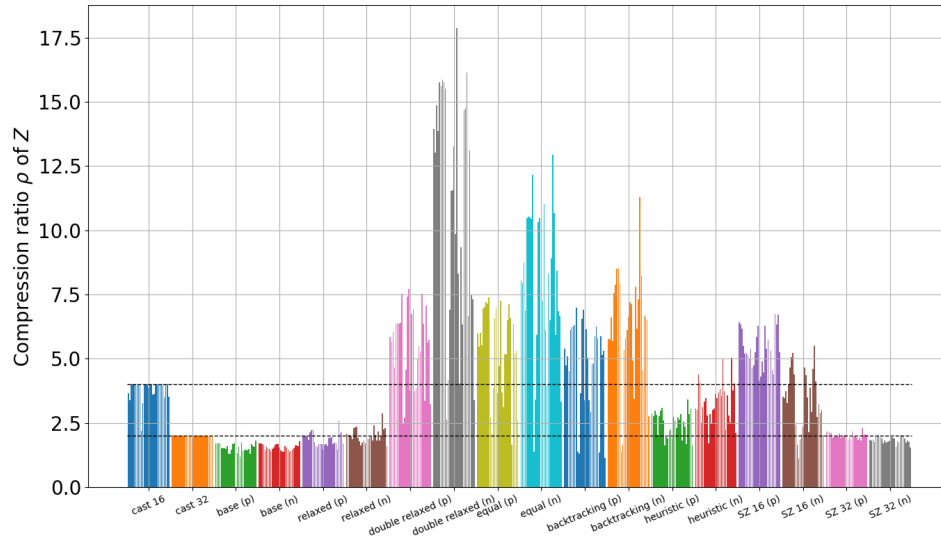


Figure 8: The compression ratio  $\rho$  of  $Z$  grouped per compression strategy for the solution of the scaled systems. Each bar per strategy corresponds to a different matrix according to its id in [Table 1](#). Horizontal reference lines are added at  $\rho = 2$  and 4.

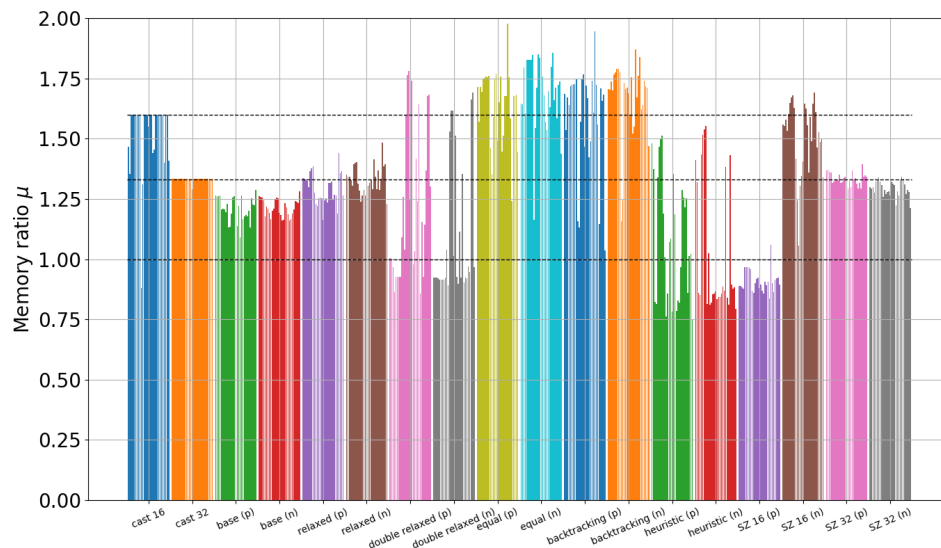


Figure 9: The memory ratio  $\mu$  grouped per compression strategy for the solution of the scaled systems. Each bar per strategy corresponds to a different matrix according to its id in [Table 1](#). Horizontal reference lines are added at  $\mu = 1$ , 1.33, and 1.6.

$\beta$	$\gamma$	dim info	iter	$\rho$	$\mu$
-100	10	N	11	8.09	1.63
-100	10	Y	11	11.05	1.68
-100	50	N	11	9.16	1.65
-100	50	Y	11	11.13	1.68
-100	100	N	11	9.20	1.65
-100	100	Y	11	11.20	1.68
-100	500	N	10	11.40	1.84
-100	500	Y	10	12.24	1.85
-100	1000	N	11	11.10	1.68
-100	1000	Y	11	11.35	1.68
-50	10	N	11	8.05	1.63
-50	10	Y	11	11.10	1.68
-50	50	N	11	8.96	1.65
-50	50	Y	11	11.11	1.68
-50	100	N	11	9.20	1.65
-50	100	Y	11	11.20	1.68
-50	500	N	10	11.39	1.84
-50	500	Y	10	12.09	1.85
-50	1000	N	11	11.04	1.68
-50	1000	Y	11	11.36	1.68
-10	10	N	11	8.08	1.63
-10	10	Y	11	11.11	1.68
-10	50	N	11	8.93	1.65
-10	50	Y	11	11.09	1.68
-10	100	N	11	9.25	1.66
-10	100	Y	11	11.18	1.68
-10	500	N	10	11.37	1.84
-10	500	Y	10	12.34	1.85
-10	1000	N	11	11.13	1.68
-10	1000	Y	11	11.40	1.68
10	10	N	11	8.08	1.63
10	10	Y	11	11.11	1.68
10	50	N	11	8.95	1.65
10	50	Y	11	11.15	1.68
10	100	N	11	9.23	1.66
10	100	Y	11	11.19	1.68
10	500	N	10	11.40	1.84
10	500	Y	10	12.35	1.85
10	1000	N	11	11.09	1.68
10	1000	Y	11	11.35	1.68
50	10	N	11	8.18	1.64
50	10	Y	11	11.09	1.68
50	50	N	11	9.16	1.65
50	50	Y	11	11.20	1.68
50	100	N	11	9.25	1.66
50	100	Y	11	11.22	1.68
50	500	N	10	11.42	1.84
50	500	Y	10	12.31	1.85
50	1000	N	11	11.07	1.68
50	1000	Y	11	11.39	1.68
100	10	N	11	8.05	1.63
100	10	Y	11	11.11	1.68
100	50	N	11	9.01	1.65
100	50	Y	11	11.14	1.68
100	100	N	11	9.26	1.66
100	100	Y	11	11.21	1.68
100	500	N	10	11.40	1.84
100	500	Y	10	12.23	1.85
100	1000	N	11	11.11	1.68
100	1000	Y	11	11.36	1.68

Table 49: Results from the scaling experiment on the scaled system.



**RESEARCH CENTRE  
BORDEAUX – SUD-OUEST**

200 avenue de la Vielle Tour  
33405 Talence Cedex

Publisher  
Inria  
Domaine de Voluceau - Rocquencourt  
BP 105 - 78153 Le Chesnay Cedex  
[inria.fr](http://inria.fr)

ISSN 0249-6399