JOURNAL OF APPLIED CRYSTALLOGRAPHY

ISSN 1600-5767

SHARP: a distributed GPU-based ptychographic solver

Stefano Marchesini,[a]* Hari Krishnan,[a] Benedikt J. Daurer,[b] David A. Shapiro,[a] Talita Perciano,[a] James A. Sethian[a] and Filipe R. N. C. Maia[b]

[a]Lawrence Berkeley National Laboratory, Berkeley, CA, USA, and [b]Uppsala University, Uppsala, Sweden. *Correspondence e-mail: smarchesini@lbl.gov

Received 28 January 2016
Accepted 17 May 2016

Edited by J. Hajdu, Uppsala University, Sweden

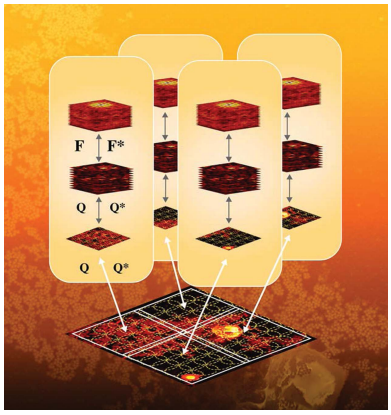
Keywords: coherent X-ray diffractive imaging; ptychography; nanoscience; X-ray microscopy; phase-contrast X-ray imaging.

Ever brighter light sources, fast parallel detectors and advances in phase retrieval methods have made ptychography a practical and popular imaging technique. Compared to previous techniques, ptychography provides superior robustness and resolution at the expense of more advanced and time-consuming data analysis. By taking advantage of massively parallel architectures, high-throughput processing can expedite this analysis and provide microscopists with immediate feedback. These advances allow real-time imaging at wavelength-limited resolution, coupled with a large field of view. This article describes a set of algorithmic and computational methodologies used at the Advanced Light Source and US Department of Energy light sources. These are packaged as a CUDA-based software environment named *SHARP* (http://camera.lbl.gov/sharp), aimed at providing state-of-the-art high-throughput ptychography reconstructions for the coming era of diffraction-limited light sources.

1. Introduction

Reconstructing the three-dimensional map of the scattering potential of a sample from measurements of its far-field scattering patterns is an important problem. It arises in a variety of fields, including optics (Fienup, 1982; Luke *et al.*, 2002), astronomy (Fienup *et al.*, 1993), X-ray crystallography (Eckert, 2012), tomography (Momose *et al.*, 1996), holography (Collier *et al.*, 1971; Marchesini *et al.*, 2008) and electron microscopy (Hawkes & Spence, 2007). As such it has been a subject of study for applied mathematicians for over a century. The fundamental problem consists of finding the correct phases that go along with the measured intensities, such that together they can be Fourier transformed into the real-space image of the sample. To help recover the correct phases from intensity measurements, a range of experimental techniques have been proposed over the years, such as interferometry/holography (Collier *et al.*, 1971), random phase masks (Marchesini *et al.*, 2008; Fannjiang & Liao, 2012; Wang & Xu, 2013) and gratings (Pfeiffer *et al.*, 2006). A variety of numerical techniques have also been recently developed, for example by approximating the problem as a matrix completion problem (Candes *et al.*, 2013) or by other convex relaxations (Waldspurger *et al.*, 2013) tractable by semidefinite programming.

Since its first demonstration (Miao *et al.*, 1999), progress has been made in solving the phase problem for a single diffraction pattern recorded from a nonperiodic object, including the dynamic update of the support (Marchesini *et al.*, 2003) and a variety of projection algortihms (Bauschke *et al.*, 2002; Marchesini, 2007*a*,*b*). Such methods, referred to as coherent diffractive imaging, attempt to recover the complete complex-valued scattering potential or electron density and the



complex exit wavefront scattered from the object, providing phase contrast as well as a way to overcome depth-of-focus limitations of regular optical systems.

Ptychography, a relatively recent technique, provides the unprecedented capability of imaging macroscopic specimens in three dimensions and attaining wavelength-limited resolution along with chemical specificity (Rodenburg, 2008). Ptychography was proposed in 1969 (Hoppe, 1969; Hegerl & Hoppe, 1970) and later experimentally demonstrated (Nellist *et al.*, 1995; Chapman, 1996), with the aim of improving the resolution in X-ray and electron microscopy. Since then it has been used in a large array of applications and shown to be a remarkably robust technique for the characterization of nanomaterials. A few software implementations of the reconstruction algorithm exist, such as *ptypy* (http://ptycho.github.io/ptypy/) and *PtychoLib* (Nashed *et al.*, 2014), and a repository for sharing experimental data has been established (Maia, 2012).

Ptychography can be used to obtain large high-resolution images. It combines the large field of view of a scanning transmission microscope with the resolution of scattering measurements. In a scanning transmission microscope operated in transmission mode, a focused beam is rastered across a sample, and the total transmitted intensity is recorded for each beam position. The pixel positions of the image obtained correspond to the beam positions used during the scan, and the value of the pixel to the intensity transmitted at that position. This limits the resolution of the image to the size of the impinging beam, which is typically limited by the quality of focusing optics and work distance constraints. In ptychography, instead of only using the total transmitted intensity, one typically records the distribution of that intensity in the far field, *i.e.* the scattering pattern produced by the interaction of the illumination with the sample. The diffracted signal contains information about features much smaller than the size of the X-ray beam, making it possible to achieve higher resolutions than with scanning techniques. The downside of having to use the intensities is that one now has to retrieve the corresponding phases to be able to reconstruct an image of the sample, which is made even more challenging by the presence of noise, experimental uncertainties and perturbations of the experimental geometry. While it is a difficult problem, it is usually tractable by making use of the redundancy inherent in obtaining diffraction patterns from overlapping regions of the sample. This redundancy also permits the technique to overcome the lack of several experimental parameters and measurement uncertainties. For example, there are methods to recover unknown illuminations (Thibault *et al.*, 2008, 2009; Hesse *et al.*, 2015; Marchesini & Wu, 2014). As a testament to their success these methods are even used as a way of characterizing high-quality X-ray optics (Kewish *et al.*, 2010; Hönig *et al.*, 2011; Guizar-Sicairos *et al.*, 2011), the wavefront of X-ray lasers (Schropp *et al.*, 2013) and EUV lithography tools (Wojdyla *et al.*, 2014).

Ptychographical phasing is a nonlinear optimization problem (Guizar-Sicairos & Fienup, 2008) still containing many open questions (Marchesini *et al.*, 2015). Several strategies, such as alternating directions (Wen *et al.*, 2012), projections, gradient (Guizar-Sicairos & Fienup, 2008), conjugate gradient, Newton (Yang *et al.*, 2011; Thibault & Guizar-Sicairos, 2012; Qian *et al.*, 2014), spectral methods (Marchesini *et al.*, 2013, 2015) and Monte Carlo (Maiden *et al.*, 2012), have been proposed to handle situations when both sample and positions (Maiden *et al.*, 2012; Beckers *et al.*, 2013; Guizar-Sicairos & Fienup, 2008; Marchesini *et al.*, 2013) are unknown parameters in high dimensions, and to handle experimental situations such as accounting for noise variance (Thibault & Guizar-Sicairos, 2012; Godard *et al.*, 2012), partial coherence (Fienup *et al.*, 1993; Abbey *et al.*, 2008; Jesse & Andrew, 2011; Whitehead *et al.*, 2009; Marchesini *et al.*, 2013; Tian *et al.*, 2014), background (Thurman & Fienup, 2009; Guizar-Sicairos & Fienup, 2009; Yang *et al.*, 2011; Marchesini *et al.*, 2013) or vibrations.

Here, we describe an algorithm approach and software environment, *SHARP* (scalable hetereogeneous adaptive real-time ptychography), that enables high-throughput streaming analysis using computationally efficient phase retrieval algorithms. The high-performance computational backend, written in C/CUDA and implemented for NVIDIA GPU architectures, is hidden from the microscopist but can be accessed and adapted to particular needs by using a Python interface or by modifying the source code.

Using *SHARP* we have built an intuitive graphical user interface that provides visual feedback on both the recorded diffraction data and the reconstructed images, throughout the data acquisition and reconstruction processes at the Advanced Light Source (ALS).

We use a mathematical formulation of ptychography which was first introduced by Yang *et al.* (2011), Qian *et al.* (2014), Wen *et al.* (2012) and Marchesini *et al.* (2013, 2015).

2. *SHARP* software environment

2.1. Forward model

In a ptychography experiment (see Fig. 1), one performs a series of diffraction measurement as a sample is rastered across an X-ray, electron or visible light beam. The illumina-

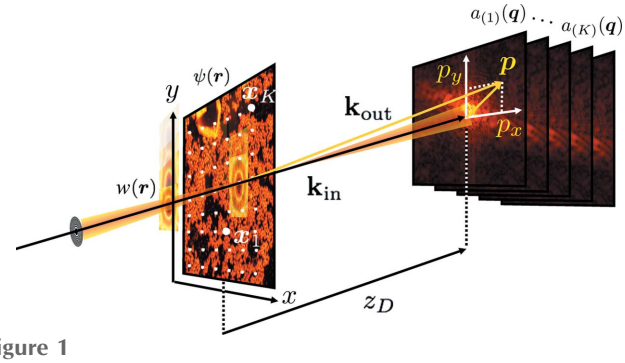


Figure 1
Experimental geometry in ptychography: an unknown sample with transmission $\psi(\mathbf{r})$ is rastered through an illuminating beam $w(\mathbf{r})$, and a sequence of diffraction measurements $\mathcal{I}_{(i)} = |a_{(i)}(\mathbf{q})|^2$ are recorded on an area detector with pixel coordinates $\mathbf{p}$ at a distance $z_{\rm D}$ from the sample.

tion is formed by an X-ray optic such as a zone plate. The measurement is performed by briefly exposing an area detector such as a CCD which records the scattered photons.

In a discrete setting, a two-dimensional small beam with distribution $w(\mathbf{r})$ of dimension $m_x \times m_y$ illuminates a subregion positioned at $\mathbf{x}_{(i)}$ (referred to as a frame) of an unknown object of interest $\psi(\mathbf{r})$ of dimension $n_x \times n_y$. Here $0 < m < n$, $i = 1, \ldots, K$ and K is the total number of frames (also referred to as 'views' in the literature). For simplicity we consider square matrices. Generalization to nonsquare matrices is straightforward but requires more indices and complicates notation.

The pixel coordinates on a detector placed at a distance $z_{\rm D}$ from the sample are described as $\mathbf{p} = (p_x, p_y, z_{\rm D})$. Under far-field and paraxial approximations the pixel coordinates are related to reciprocal space coordinates as

$$\begin{aligned}\mathbf{q} = \mathbf{k}_{\rm out} - \mathbf{k}_{\rm in} &= \frac{1}{\lambda}\left[\frac{(p_x, p_y, z_{\rm D})}{(p_x^2 + p_y^2 + z_{\rm D}^2)^{1/2}} - (0, 0, 1)\right] \\ &\simeq \frac{1}{\lambda z_{\rm D}}\left(p_x, p_y, 0\right),\end{aligned} \tag{1}$$

where $\mathbf{k}_{\rm in} = (0, 0, k)$ and $\mathbf{k}_{\rm out} = k\mathbf{p}/|\mathbf{p}|$ are the incident and scattered wavevectors that satisfy $|\mathbf{k}_{\rm in}| = |\mathbf{k}_{\rm out}| = k = 1/\lambda$, and λ is the wavelength. With a distance p_m from the center to the edge of the detector, the diffraction-limited resolution (half-period) of the microscope is given by the length scale $r = \lambda z_{\rm D}/(2p_m)$. As a consequence, the coordinates in reciprocal and real space are defined as

$$\mathbf{q} = \left(\frac{\mu}{mr}, \frac{\nu}{mr}\right), \quad \mu, \nu \in \{0, \ldots, m-1\} \tag{2}$$

and

$$\mathbf{r} = (r\mu, r\nu), \quad \mu, \nu \in \{0, \ldots, m-1\}, \tag{3}$$

$$\mathbf{x}_{(i)} = (r\mu', r\nu'), \quad \mu', \nu' \in \{0, \ldots, n-m\}. \tag{4}$$

While $\mathbf{x}_{(i)}$ is typically rastered on a coarser grid, $\mathbf{r} + \mathbf{x}_{(i)}$ spans a finer grid of dimension $n \times n$.

In other words, we assume that a sequence of K diffraction intensity patterns $\mathcal{I}_{(i)}(\mathbf{q})$ are collected as the position of the object is rastered on the position $\mathbf{x}_{(i)}$. The simple transform $a_{(i)} = [\mathcal{I}_{(i)}(\mathbf{q})]^{1/2}$ is a variance-stabilizing transform for Poisson noise (Anscombe, 1948; Mäkitalo & Foi, 2013). The relationship between the amplitude $a_{(i)}(\mathbf{q})$, the illumination function $w(\mathbf{r})$ and an unknown object $\psi(\mathbf{r})$ to be estimated can be expressed as follows:

$$a_{(i)}(\mathbf{q}) = \left|\mathcal{F} w(\mathbf{r})\psi(\mathbf{r} + \mathbf{x}_{(i)})\right|. \tag{5}$$

$\mathcal{F}$ is the two-dimensional discrete Fourier transform,

$$(\mathcal{F}f)(\mathbf{q}) = \frac{1}{(m^2)^{1/2}} \sum_{\mathbf{r}} \exp(2\pi i \mathbf{q}\cdot\mathbf{r}) f(\mathbf{r}), \tag{6}$$

where the sum over $\mathbf{r}$ is given on all the indices $m \times m$ of $\mathbf{r}$. We define an operator $T_{(i)}$ that extracts a frame out of an image ψ and build the illumination operator $\mathbf{Q}_{(i)}$, which scales the extracted frame pointwise by the illumination function w:

$$\mathbf{Q}_{(i)}[\psi](\mathbf{r}) = w(\mathbf{r})\psi(\mathbf{r} + \mathbf{x}_{(i)}) = w(\mathbf{r})T_{(i)}[\psi](\mathbf{r}) \ = z_{(i)}(\mathbf{r}). \tag{7}$$

With the operator $\mathbf{Q}$, equation (5) can be represented compactly as

$$\mathbf{a} = |\mathbf{F}\mathbf{Q}\psi^{\vee}| \quad \text{or} \quad \begin{cases} \mathbf{a} = |\mathbf{F}\mathbf{z}| \\ \mathbf{z} = \mathbf{Q}\psi^{\vee} \end{cases} \tag{8}$$

where $\psi^{\vee}$ denotes the linearized version of the image (the superscript will be omitted for simplicity). More explicitly,

$$\overbrace{\begin{bmatrix} a_{(1)} \\ \vdots \\ a_{(K)} \end{bmatrix}}^{\mathbf{a}\in\mathbb{R}^{Km^2}} = \left| \overbrace{\begin{bmatrix} F & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & F \end{bmatrix}}^{\mathbf{F}\in\mathbb{C}^{Km^2\times Km^2}} \overbrace{\begin{bmatrix} z_{(1)} \\ \vdots \\ z_{(K)} \end{bmatrix}}^{\mathbf{z}\in\mathbb{C}^{Km^2}} \right|, \tag{9}$$

$$\overbrace{\begin{bmatrix} z_{(1)} \\ \vdots \\ z_{(K)} \end{bmatrix}}^{\mathbf{z}\in\mathbb{C}^{Km^2}} = \overbrace{\begin{bmatrix} \mathrm{diag}(w)T_{(1)} \\ \vdots \\ \mathrm{diag}(w)T_{(K)} \end{bmatrix}}^{\mathbf{Q}\in\mathbb{C}^{Km^2\times n^2}} \overbrace{\begin{bmatrix} \psi_1 \\ \vdots \\ \psi_{n^2} \end{bmatrix}}^{\psi\in\mathbb{C}^{n^2}}. \tag{10}$$

Here $\mathbf{z}$ are K frames extracted from the object ψ and multiplied by the illumination function w, and $\mathbf{F}$ is the associated two-dimensional discrete Fourier transform matrix when we write everything in the stacked form (Marchesini *et al.*, 2013). When both the sample and the illumination are unknown, we can express the relationship (10) between the image ψ, the illumination w and the frames $\mathbf{z}$ in two forms:

$$\mathbf{z} = \mathbf{Q}\psi = \mathrm{diag}(\mathbf{S}w)\mathbf{T}\psi = \mathrm{diag}(\mathbf{T}\psi)\mathbf{S}w. \tag{11}$$

$\mathbf{S} \in \mathbb{R}^{Km^2\times m^2}$ denotes the operator that replicates the illumination w into a stack of K frames, since $\mathbf{Q}\psi = \mathrm{diag}(\mathbf{S}w)\mathbf{T}\psi$ is the entry-wise product of $\mathbf{T}\psi$ and $\mathbf{S}w$. Equation (11) can be used to find ψ or w from $\mathbf{z}$ and the other variable.

The Fourier transform relationship used in equations (5), (8) and (9) is valid under the far-field and paraxial approximation, which is the focus of the current release of *SHARP*. For experimental geometries such as near-field (Stockmar *et al.*, 2013), Fresnel (Vine *et al.*, 2009), Fourier ptychography (Zheng *et al.*, 2013), through-focus (Marrison *et al.*, 2013), undersampled (Edo *et al.*, 2013) and partially coherent multiplexed geometries (Tian *et al.*, 2014; Batey *et al.*, 2014; Dong *et al.*, 2014), and to account for noise variance, one can replace the simple Fourier transform with the appropriate propagator (Marchesini *et al.*, 2013) and variance stabilization (Yang *et al.*, 2011).

2.2. Phase retrieval

Projection operators form the basis of every iterative projection, and projected gradient algorithms are implemented in *SHARP* and accessible through a library. The projection $P_{\mathbf{a}}$ ensures that the frames $\mathbf{z}$ match the experiment, that is, they satisfy equation (9), and is referred to as the data projector:

$$P_{\mathbf{a}}\mathbf{z} = \mathbf{F}^{*}\frac{\mathbf{Fz}}{|\,\mathbf{Fz}\,|}\mathbf{a}. \tag{12}$$

The projection $P_{\mathbf{Q}}$ onto the range of $\mathbf{Q}$ (see Fig. 2),

$$P_{\mathbf{Q}} = \mathbf{Q}(\mathbf{Q}^{*}\mathbf{Q})^{-1}\mathbf{Q}^{*}, \tag{13}$$

ensures that overlapping frames $\mathbf{z}$ are consistent with each other and satisfy equation (10).

The projector $P_{\mathbf{a}}$ is relatively robust to Poisson noise (Anscombe, 1948), but weighting factors to account for noisy pixels can be easily added (Qian *et al.*, 2014).

Using relationship (11), we can update the image ψ from w and frames $\mathbf{z}$:

$$\psi \leftarrow \frac{\mathbf{Q}^{*}\mathbf{z}}{\mathbf{Q}^{*}\mathbf{Q}}. \tag{14}$$

Or, we can update the illumination w from an image ψ and frames $\mathbf{z}$ (Thibault *et al.*, 2008, 2009), multiplying equation (11) on the left by $\mathrm{diag}\mathbf{T}\overline{\psi}$ and solving for w:

$$w \leftarrow \frac{\mathbf{S}^{*}\mathrm{diag}(\mathbf{T}\overline{\psi})\mathbf{z}}{\mathbf{S}^{*}\mathbf{T}|\psi|^{2}}, \tag{15}$$

where $\overline{\psi}$ denotes the complex conjugate of ψ. See Marchesini & Wu (2014) for alternative updates, and Hesse *et al.* (2015) for the convergence theory behind a similar blockwise optimization strategy. Several possible pathologies need to be accounted for when updating both ψ and w:

(*a*) Combined drift of the illumination and the image in real space is eliminated by keeping the illumination in the center of the frame by computing the center of mass and correcting for drifts after every update of the illumination.

(*b*) Fourier space drifts and grid pathologies are suppressed by enforcing either the absolute value $a_w = |\mathcal{F}w_0|$ or support m_w of the Fourier transform of the unknown illumination w_0.

(*c*) A possible global phase factor between the solution and the reconstruction is taken into account in the error calculation.

A typical reconstruction with *SHARP* uses the following sequence:

(1) Input data $\mathcal{I}(\mathbf{q})$, translations $\mathbf{x}$. Optional inputs: initial image $\psi^{(0)}$, illumination $w^{(0)}$, illumination Fourier mask m_w and illumination Fourier amplitudes a_w.

(2) If $w^{(0)}$ is not provided, initialize illumination by setting $w^{(0)}$ to the inverse Fourier transform of the square root of the average frame.

(3) If $\psi^{(0)}$ is not provided, initialize the image by filling $\psi^{(0)}$ with random numbers uniformly drawn from [0, 1).

(4) Build up $\mathbf{Q}$, $\mathbf{Q}^{*}$ and $(\mathbf{Q}^{*}\mathbf{Q})^{-1}$, and frames $\mathbf{z}^{(0)} = \mathbf{Q}\psi^{(0)}$.

(5) Update the frames $\mathbf{z}$ according to Luke (2005) using the projector operators defined in equations (12) and (13):

$$\mathbf{z}^{(l)} := [2\beta P_{\mathbf{Q}}P_{\mathbf{a}} + (1-2\beta)P_{\mathbf{a}} + \beta(P_{\mathbf{Q}} - I)]\mathbf{z}^{(l-1)}, \tag{16}$$

where $\beta \in (0.5, 1]$ is a scalar value set by the user (set to 0.75 by default, which works in most cases).

(6) Update image $\psi^{(\ell)}$ using equation (14).

(7) If desired, compute a new illumination w using equation (15). If m_w is given apply the illumination Fourier mask constraint:

$$w^{(\ell)} := \mathcal{F}^{-1}\{(\mathcal{F}w)m_w\}. \tag{17}$$

Else if w_I is given apply the illumination Fourier intensities constraint:

$$w^{(\ell)} := \mathcal{F}^{-1}\left\{\frac{\mathcal{F}w}{|\mathcal{F}w|}a_w\right\}. \tag{18}$$

Else simply keep the unconstrained illumination $w^{(\ell)} := w$. Now compute the center of mass of $w^{(\ell)}$ and shift it to fix the translation of the object.

(8) If desired perform background retrieval, that is, estimate static background and remove it in the iteration as described by Marchesini *et al.* (2013, p. 7, equation 30).

(9) Iterate from 5 until one of the metrics from equations (19)–(22) drops below a user-defined level or until a maximum iteration for time-critical applications, and return $\psi^{(\ell)}$ and w.

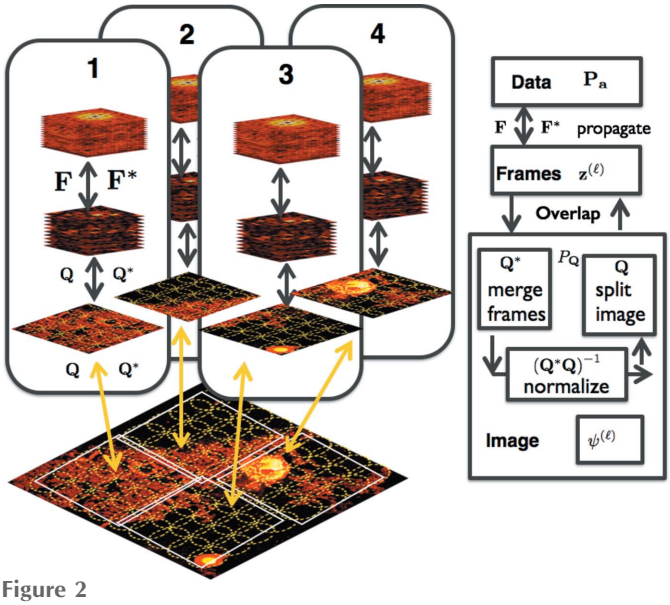


Figure 2
Schematic of the ptychographic reconstruction algorithm implemented in *SHARP*. The iterative reconstruction scheme is shown on the right. To achieve the highest possible throughput and scalability one has to parallelize across multiple GPUs, as shown on the left for the case of four GPUs. As most ptychographic scans use a constant density of scan point across the object, we expect to be able to achieve a very even division, resulting in good load balancing. *SHARP* enforces an overlap constraint between the images produced by each of the GPUs, and also enforces that the illuminations recovered on each GPU agree with each other. This is done by default at every iteration.

The metrics $\varepsilon_{\mathbf{a}}$, $\varepsilon_{\mathbf{Q}}$, ε_{Δ} used to monitor progress are the normalized mean square root error from the corresponding projections $\mathbf{z}$:

$$\varepsilon_{\mathbf{a}}(\mathbf{z}) = \frac{\|(P_{\mathbf{a}} - I)\mathbf{z}\|}{\|\mathbf{a}\|}, \tag{19}$$

$$\varepsilon_{\mathbf{Q}}(\mathbf{z}) = \frac{\|(P_{\mathbf{Q}} - I)\mathbf{z}\|}{\|\mathbf{a}\|}, \tag{20}$$

$$\varepsilon_{\Delta}(\mathbf{z}^{(l)}, \mathbf{z}^{(l-1)}) = \frac{\|\mathbf{z}^{(l)} - \mathbf{z}^{(l-1)}\|}{\|\mathbf{a}\|}, \tag{21}$$

where I is the identity operator and $\mathbf{z}^{(0)} = 0$.

For benchmarking purposes, when using a simulation from a known solution ψ_0, the following metric can also be used:

$$\varepsilon_0(\mathbf{z}) = \frac{1}{\|\mathbf{Q}^*\mathbf{z}_0\|} \min_{\varphi} \|\mathbf{Q}^*[\exp(i\varphi)\mathbf{z} - \mathbf{z}_0]\|, \tag{22}$$

where φ is an arbitrary global phase factor, and $\mathbf{z}_0 = \mathbf{Q}\psi_0$. Notice the additional scaling factor $\mathbf{Q}^*$ used in ε_0.

The initial values for the input data and translations can either be loaded from file or set by a Python interface. The starting 'zeroth' initial image is loaded from file, set to a random image or taken as a constant image.

2.3. Computational methodology

SHARP was developed to achieve the highest performance, taking advantage of the algorithm described earlier and using a distributed computational backend. The ptychographic reconstruction algorithm requires one to compute the product of several linear operators ($\mathbf{Q}$, $\mathbf{Q}^*$, $\mathbf{F}$, $\mathbf{F}^*$, $\mathbf{S}$, $\mathbf{S}^*$) on a set of frames $\mathbf{z}$, an image ψ and an illumination w several times. We use a distributed GPU architecture across multiple nodes for this task (Fig. 2).

To implement fast operators, a set of GPU kernels and MPI communication are necessary. The split ($\mathbf{Q}\psi$) and overlap ($\mathbf{Q}^*\mathbf{z}$) kernels are among the most bandwidth demanding kernels and play an important role in the process.

The strategy used to implement those kernels impacts directly on the overall performance of the reconstruction algorithm. To divide the problem among multiple nodes, *SHARP* initially determines the size of the final image from the list of translations, frames size and resolution. It subsequently assigns a list of translations to every node and loads the corresponding frames onto GPUs.

The split ($\mathbf{Q}\psi$) and fast Fourier transform ($\mathbf{F}$) operations are easily parallelized because of the framewise intrinsic independence. Summing the frames onto an image ($\mathbf{Q}^*\mathbf{z}$) requires a reduction for every image pixel across neighboring MPI nodes. Within each GPU the image is divided into blocks, and we first determine which frames contribute to each block. The contributing frames are summed and then the resulting image is summed across all MPI nodes. We use shared memory or constant memory, depending on GPU compute capability, to store frame translations, and we use kernel fusion to reduce access to global memory. The last step of summing across all MPI nodes does not necessarily have to be done at every iteration, at the cost of slower convergence (Liu *et al.*, 2015), but that is the default.

The timing to compute the overlap at each iteration depends on the size of the image and the number of frames on top of each pixel, *i.e.* the density but not the size of the frames.

In addition to the high performance ptychographic algorithm, the *SHARP* software environment provides a flexible and modular framework which can be changed and adapted to different needs. Furthermore, the user has control of several options for the reconstruction algorithm, which can be used to guarantee a balance between performance and quality of the results. These include the choice of illumination Fourier mask, illumination Fourier intensities and the β parameter, as well as how often to perform different operations such as illumination retrieval, background retrieval and synchronization of the different GPUs. For more details we refer the reader to the documentation (http://www.camera.lbl.gov/sharp).

3. Applications and performance

SHARP enables high-throughput streaming analysis using computationally efficient phase retrieval algorithms. In this section we describe a typical dataset and sample that can be collected in less than 1 min at the ALS, and the computational backend to provide fast feedback to the microscopist.

To characterize our performance, we use both simulations and experimental data. We use simulations to compare the convergence of the reconstruction algorithm with the 'true solution' and characterize the effect of different light sources, contrast, scale, noise, detectors or samples for which no data exist yet.

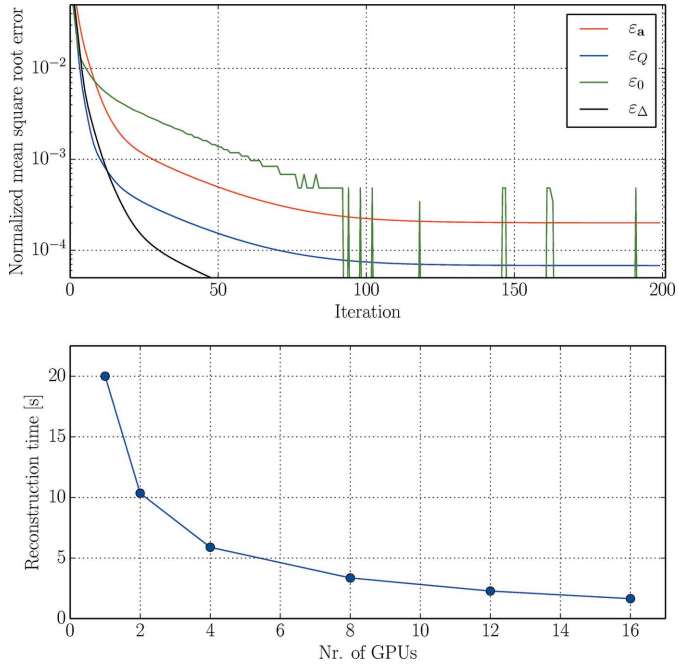


Figure 3
Convergence rate (top) per iteration and timing (bottom) to process 10 000 frames of dimension 128 × 128 extracted from an image of size 1000 × 1000 as a function of the number of nodes. All residuals decrease rapidly; numerical precision limits the (weighted) comparison with the known solution $\varepsilon'(\mathbf{z})$. Reconstruction is achieved ($\varepsilon_0 < 5e-4$) in under 2 s using a cluster with four compute nodes with four GTX Titan GPUs per node (16 total, 43 000 cores), 96 GB GPU memory, 1 TB RAM and 24 TB storage, infiniband. The timing contributions for the corresponding computational kernels are $(\mathbf{Q}^*\mathbf{Q})^{-1}\mathbf{Q}^*$ 30%, $\mathbf{F}$, $\mathbf{F}^*$ 20%, $\mathbf{Q}$ 20%, $\mathbf{S}^*$ 5%, elementwise operations 20%, and residual calculation 5%. No illumination retrieval was done, as the exact illumination was given. The simulation was done using periodic boundary conditions to avoid edge effects.

Experimental data from ALS used to characterize battery materials, green cement and magnetic materials at different wavelengths and orientations have been successfully reconstructed (Shapiro *et al.*, 2014; Yu *et al.*, 2015; Bae *et al.*, 2015; Li *et al.*, 2015; Shi *et al.*, 2016) using the software described in this article.

We also describe a streaming example in which a frontend that operates very close to the actual experiment sends the data to the reconstruction backend which runs remotely on a GPU/CPU cluster. Further details about the streaming frontend and processing backend pipeline will be published in an upcoming paper by the same authors.

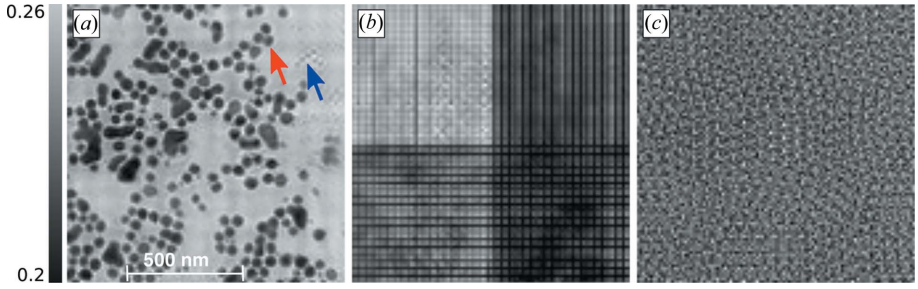


Figure 4
Reconstruction of a test sample consisting of gold balls with diameters of 50 and 10 nm. Detector pixel size 30 µ, 1920 × 960 pixels 80 mm downstream from the sample, cropped and downsampled to 128, scan of 50 × 50 points; illumination is generated by a zone plate with a diameter of 220 µm and 60 nm outer zone width. (*a*) Phase image generated by *SHARP* using the algorithm described in §2.2, applying the illumination Fourier mask constraint and turning on background retrieval. The red arrow points to a collection of 50 nm balls, while the blue arrow points to a collection of 10 nm balls. The pixel size is 10 nm. (*b*) Same as (*a*) except without enforcing the illumination Fourier mask. (*c*) Same as (*c*) but without using the background retrieval algorithm.

3.1. Simulations and performance

As a demonstration, we start from a sample that was composed of colloidal gold nanoparticles of 50 and 10 nm deposited on a transparent silicon nitride membrane. An experimental image was obtained by scanning electron microscopy, which provides high resolution and contrast but can only view the surface of the sample.

We simulate a complex transmission function by scaling the image amplitude from 0 to 50 nm thickness, and using the complex index of refraction of gold at 750 eV energy from http://henke.lbl.gov. The illumination is generated by a zone plate with a diameter of 220 µm and 60 nm outer zone width, discretized into 128 × 128 pixels in the far field.

See Fig. 3 for a performance example.

3.2. Experimental example

Fig. 4 shows ptychographic reconstructions of a dataset generated from a sample consisting of gold balls with diameters of 50 and 10 nm. The data were generated using 750 eV X-rays at beamline 5.3.2.1 of the ALS, with high-stability position control of the soft X-ray scanning transmission microscope. The exposure time was 1 s and the dataset consists of a square scan grid with 40 nm spacing [see Shapiro *et al.* (2014) for details of the experimental setup]. The reconstructions consisted of 300 iterations of the RAAR algorithm with an illumination retrieval and background retrieval step every other iteration. The initial illumination is generated by (1) computing the average of the measurements, (2) setting everything below a threshold to 0 and everything above that threshold to a constant value, and (3) applying an inverse fast Fourier transform. The image is initialized with complex independent identically distributed pixels.

3.3. Interface and streaming

Common processing pipelines used for ptychographic experiments usually have a series of I/O operations and many different components involved. We have developed a streaming pipeline, to be deployed at the COSMIC beamline at the ALS, which allows users to monitor and quickly act upon changes along the experimental and computational pipeline.

The streaming pipeline is composed of a frontend and a backend (Fig. 5). The frontend consists of a graphical user interface (see Fig. 5), a worker that grabs frames from the detector, and an interface that monitors network activity and experimental parameters (position, wavelength, exposure *etc.*) and provides a live view of the ongoing reconstruction.

On the backend side, the streaming infrastructure is composed of a communication handler and a collection of workers addressing different tasks such as dark calibration, detector correction, data reduction, ptychographic reconstruction and writing output to file.

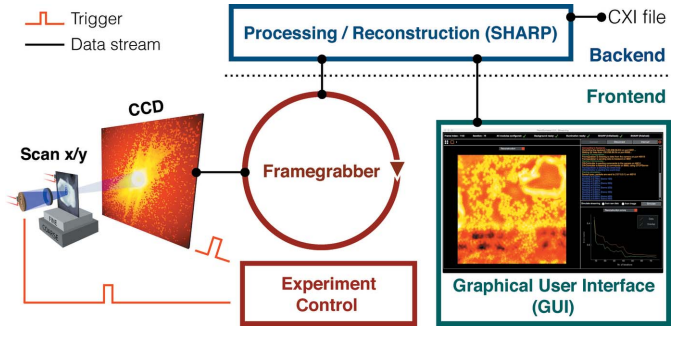


Figure 5
Overview of the components involved in the software structure of the streaming pipeline. In order to maximize the performance of this streaming framework, the frontend operates very close to the actual experiment, while the backend runs remotely on a powerful GPU/CPU cluster. As soon as diffraction data are recorded by the CCD camera, a live view of the ptychographic reconstruction is transmitted to the graphical user interface, and the user is able to control and monitor (top panel) the current status of the data streams and analysis (bottom right panel).

This software architecture allows users an intuitive, flexible and responsive monitoring and control of their experiments. Such a tight integration between data aquisition and analysis is required to give users the feedback they expect from a STXM instrument.

4. Conclusions

In this paper we described *SHARP*, a high-performance software environment for ptychography reconstructions, and its application as part of the quick feedback system used by the ptychographic microscopes installed at the ALS.

Our software provides a modular interface to the high-performance computational backend and can be adapted to different needs. Its fast throughput provides near-real-time feedback to microscopists, and this also makes it suitable as a corner stone for demanding higher-dimensional analysis such as spectro-ptychography or tomo-ptychography.

With the coming new-generation light sources and faster detectors, the ability to quickly analyze vast quantities of data to obtain large high-dimensional images will be an enabling tool for science.

Acknowledgements

We acknowledge useful discussions with Chao Yang, H.-T. Wu, J. Qian and Z. Wen. This work was partially funded by the Center for Applied Mathematics for Energy Research Applications, a joint ASCR-BES funded project within the Office of Science, US Department of Energy, under contract No. DOE-DE-AC03-76SF00098, by the Swedish Research Council and by the Swedish Foundation for Strategic Research.

References

Abbey, B., Nugent, K. A., Williams, G. J., Clark, J. N., Peele, A. G., Pfeifer, M. A., de Jonge, M. & McNulty, I. (2008). *Nat. Phys.* **4**, 394–398.

Anscombe, F. J. (1948). *Biometrika*, **35**, 246–254.

Bae, S., Taylor, R., Shapiro, D., Denes, P., Joseph, J., Celestre, R., Marchesini, S., Padmore, H., Tyliszczak, T., Warwick, T., Kilcoyne, D., Levitz, P. & Monteiro, P. J. M. (2015). *J. Am. Ceramic Soc.* **98**, 4090–4095.

Batey, D. J., Claus, D. & Rodenburg, J. M. (2014). *Ultramicroscopy*, **138**, 13–21.

Bauschke, H. H., Combettes, P. L. & Luke, D. R. (2002). *J. Opt. Soc. Am. A*, **19**, 1334–1345.

Beckers, M., Senkbeil, T., Gorniak, T., Giewekemeyer, K., Salditt, T. & Rosenhahn, A. (2013). *Ultramicroscopy*, **126**, 44–47.

Candes, E. J., Eldar, Y. C., Strohmer, T. & Voroninski, V. (2013). *SIAM J. Imaging Sci.* **6**, 199–225.

Chapman, H. N. (1996). *Ultramicroscopy*, **66**, 153–172.

Collier, R. J., Burckhardt, C. B. & Lin, L. H. (1971). *Optical Holography*. New York: Academic Press.

Dong, S., Shiradkar, R., Nanda, P. & Zheng, G. (2014). *Biomed. Opt. Express*, **5**, 1757–1767.

Eckert, M. (2012). *Acta Cryst.* A**68**, 30–39.

Edo, T. B., Batey, D. J., Maiden, A. M., Rau, C., Wagner, U., Pešić, Z. D., Waigh, T. A. & Rodenburg, J. M. (2013). *Phys. Rev. A*, **87**, 053850.

Fannjiang, A. & Liao, W. (2012). *J. Opt. Soc. Am. A*, **29**, 1847–1859.

Fienup, J. R. (1982). *Appl. Opt.* **21**, 2758–2769.

Fienup, J. R., Marron, J. C., Schulz, T. J. & Seldin, J. H. (1993). *Appl. Opt.* **32**, 1747–1767.

Godard, P., Allain, M., Chamard, V. & Rodenburg, J. (2012). *Opt. Express*, **20**, 25914–25934.

Guizar-Sicairos, M. & Fienup, J. R. (2008). *Opt. Express*, **16**, 7264–7278.

Guizar-Sicairos, M. & Fienup, J. R. (2009). *Opt. Express*, **17**, 2670–2685.

Guizar-Sicairos, M., Narayanan, S., Stein, A., Metzler, M., Sandy, A. R., Fienup, J. R. & Evans-Lutterodt, K. (2011). *Appl. Phys. Lett.* **98**, 111108.

Hawkes, P. W. & Spence, J. C. H. (2007). *Science of Microscopy*. New York: Springer.

Hegerl, R. & Hoppe, W. (1970). *Ber. Bunsen-Ges. Phys. Chem.* **74**, 1148.

Hesse, R., Luke, D. R., Sabach, S. & Tam, M. K. (2015). *SIAM J. Imaging Sci.* **8**, 426–457.

Hönig, S., Hoppe, R., Patommel, J., Schropp, A., Stephan, S., Schöder, S., Burghammer, M. & Schroer, C. G. (2011). *Opt. Express*, **19**, 16324–16329.

Hoppe, W. (1969). *Acta Cryst.* A**25**, 495–501.

Jesse, N. C. & Andrew, G. P. (2011). *Appl. Phys. Lett.* **99**, 154103.

Kewish, C. M., Thibault, P., Dierolf, M., Bunk, O., Menzel, A., Vila-Comamala, J., Jefimovs, K. & Pfeiffer, F. (2010). *Ultramicroscopy*, **110**, 325–329.

Li, Y., Meyer, S., Lim, J., Lee, S. C., Gent, W. E., Marchesini, S., Krishnan, H., Tyliszczak, T., Shapiro, D., Kilcoyne, A. L. D. & Chueh, W. C. (2015). *Adv. Mater.* **27**, 6590.

Liu, J., Wright, S. J., Ré, C., Bittorf, V. & Sridhar, S. (2015). *J. Machine Learning Res.* **16**, 285–322.

Luke, D. R., Burke, J. V. & Lyon, R. G. (2002). *SIAM Rev.* **44**, 169–224.

Maia, F. R. N. C. (2012). *Nat. Methods*, **9**, 854–855.

Maiden, A. M., Humphry, M. J., Sarahan, M. C., Kraus, B. & Rodenburg, J. M. (2012). *Ultramicroscopy*, **120**, 64–72.

Mäkitalo, M. & Foi, A. (2013). *IEEE Trans. Image Processing*, **22**, 91–103.

Marchesini, S. (2007*a*). *J. Opt. Soc. Am. A*, **24**, 3289–3296.

Marchesini, S. (2007*b*). *Rev. Sci. Instrum.* **78**, 011301.

Marchesini, S., Boutet, S. *et al.* (2008). *Nat. Photon.* **2**, 560–563.

Marchesini, S., He, H., Chapman, H. N., Hau-Riege1, S. P., Noy, A., Howells, M. R., Weierstall, U. & Spence, J. C. H. (2003). *Phys. Rev. B*, **68**, 140101.

Marchesini, S., Schirotzek, A., Yang, C., Wu, H.-T. & Maia, F. R. N. C. (2013). *Inverse Problems*, **29**, 115009.

Marchesini, S., Tu, Y.-C. & Wu, H.-T. (2015). *Appl. Comput. Harmonic Anal.* doi:10.1016/j.acha.2015.06.005.

Marchesini, S. & Wu, H.-T. (2014). Technical Report LBNL-6734E, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. *arXiv*:1408.1922.

Marrison, J., Räty, L., Marriott, P. & O'Toole, P. (2013). *Sci. Rep.* **3**, 02369.

Miao, J., Charalambous, P., Kirz, J. & Sayre, D. (1999). *Nature*, **400**, 342–344.

Momose, A., Takeda, T., Itai, Y. & Hirano, K. (1996). *Nat. Med.* **2**, 473–475.

Nashed, Y. S. G., Vine, D. J., Peterka, T., Deng, J., Ross, R. & Jacobsen, C. (2014). *Opt. Express*, **22**, 32082–32097.

Nellist, P. D., McCallum, B. C. & Rodenburg, J. M. (1995). *Nature*, **374**, 630–632.

Pfeiffer, F., Weitkamp, T., Bunk, O. & David, C. (2006). *Nat. Phys.* **2**, 258–261.

Qian, J., Yang, C., Schirotzek, A., Maia, F. R. N. C. & Marchesini, S. (2014). *Inverse Problems Appl. Contemp. Math.* **615**, 261–280.

Rodenburg, J. M. (2008). *Ptychography and Related Diffractive Imaging Methods*, Advances in Imaging and Electron Physics, Vol. 150, pp. 87–184. San Diego: Elsevier.

Schropp, A. *et al.* (2013). *Sci. Rep.* **3**, 01633.
Shapiro, D. A., Yu, Y.-S., Tyliszczak, T., Cabana, J., Celestre, R., Chao, W., Kaznatcheev, K., Kilcoyne, A. L. D., Maia, F., Marchesini, S., Meng, Y. S., Warwick, T., Yang, L. L. & Padmore, H. A. (2014). *Nat. Photon.* **8**, 765–769.
Shi, X., Fischer, P., Neu, V., Elefant, D., Lee, J. C. T., Shapiro, D. A., Farmand, M., Tyliszczak, T., Shiu, H.-W., Marchesini, S., Roy, S. & Kevan, S. D. (2016). *Appl. Phys. Lett.* **108**, 094103.
Stockmar, M., Cloetens, P., Zanette, I., Enders, B., Dierolf, M., Pfeiffer, F. & Thibault, P. (2013). *Sci. Rep.* **3**, 01927.
Thibault, P., Dierolf, M., Bunk, O., Menzel, A. & Pfeiffer, F. (2009). *Ultramicroscopy*, **109**, 338–343.
Thibault, P., Dierolf, M., Menzel, A., Bunk, O., David, C. & Pfeiffer, F. (2008). *Science*, **321**, 379–382.
Thibault, P. & Guizar-Sicairos, M. (2012). *New J. Phys.* **14**, 063004.
Thurman, S. T. & Fienup, J. R. (2009). *J. Opt. Soc. Am. A*, **26**, 1008–1014.
Tian, L., Li, X., Ramchandran, K. & Waller, L. (2014). *Biomed. Opt. Express*, **5**, 2376–2389.
Vine, D. J., Williams, G. J., Abbey, B., Pfeifer, M. A., Clark, J. N., De Jonge, M. D., McNulty, I., Peele, A. G. & Nugent, K. A. (2009). *Phys. Rev. A*, **80**, 063823.
Waldspurger, I., d'Aspremont, A. & Mallat, S. (2013). *Math. Program.* **149**, 47–81.
Wang, Y. & Xu, Z. (2013). *arXiv*:1310.0873.
Wen, Z., Yang, C., Liu, X. & Marchesini, S. (2012). *Inverse Problems*, **28**, 115010.
Whitehead, L. W., Williams, G. J., Quiney, H. M., Vine, D. J., Dilanian, R. A., Flewett, S., Nugent, K. A., Peele, A. G., Balaur, E. & McNulty, I. (2009). *Phys. Rev. Lett.* **103**, 243902.
Wojdyla, A., Miyakawa, R. & Naulleau, P. (2014). *Proc. SPIE*, **9048**, 904839.
Yang, C., Qian, J., Schirotzek, A., Maia, F. R. N. C. & Marchesini, S. (2011). *Iterative Algorithms for Ptychographic Phase Retrieval*. Report No. 4598E, Lawrence Berkeley National Laboratory, USA. *arXiv*:1105.5628.
Yu, Y.-S. *et al.* (2015). *Nano Lett.* **15**, 4282–4288.
Zheng, G., Horstmeyer, R. & Yang, C. (2013). *Nat. Photon.* **7**, 739–745.