High speed spatially multimode $\Lambda$-type atomic memory with arbitrary frequency detuning

T. Golubeva, Yu. Golubev and O. Mishina, A. Bramati, J. Laurat, E. Giacobino

1 St. Petersburg State University, 198504 St. Petersburg, Stary Petershof, ul. Ul’yanovskaya, 1, Russia
2 Laboratoire Kastler Brossel, Université Pierre et Marie Curie, Ecole Normale Supérieure, CNRS, Case 74, 4 place Jussieu, 75252 Paris Cedex 05, France
3 Theoretische Physik, Universität des Saarlandes, D-66123 Saarbrücken, Germany

December 22, 2011

Abstract. We present a general model for an atomic memory using ultra-short pulses of light, which allows both spatial and temporal multimode storage. The process involves the storage of a faint quantum light pulse into the spin coherence of the ground state of $\Lambda$-type 3-level atoms, in the presence of a strong driving pulse. Our model gives a full description of the evolution of the field and of the atomic coherence in space and time throughout the writing and the read-out processes. It is valid for any frequency detuning, from the resonant case to the Raman case, and allows a detailed optimization of the memory efficiency.

PACS. XX.XX.XX No PACS code given

1 Introduction

For quantum telecommunications and for quantum information processing, memory registers able to store quantum information without measuring it are essential devices. A quantum memory relies on an efficient coupling between light and matter, allowing reversible mapping of quantum photonic information in and out of the material system. In the past years, several protocols have been developed theoretically and experimentally[1,2]. Storage and retrieval of some of the basic states of light for quantum communication such as a polarization q-bit [3], squeezed light [4,5] and entangled photons [6,7] or faint coherent pulses at the level of one to few photons have been realized [8,9,10,11]. Recent experiments have achieved large efficiencies [10,11].

However, the processing speed and the available bandwidth of the memories remain a challenge for quantum memories. The first quantum memory registers proposed more than a decade ago [12] involve the transfer of quantum information from light to atoms (writing) and back from atoms to light (retrieval), using electromagnetically induced transparency (EIT) in atomic three-level transitions, and this process implies a limited bandwidth. The storage protocol relies on a strong control field, generating EIT for the weak field that carries the quantum signal to be stored. The group velocity for the signal field is strongly reduced and the signal pulse is compressed by several orders of magnitude. A signal pulse can thus be contained inside the atomic medium, and before it propagates outside the medium, the control is switched off. The quantum variables of the signal field are then converted from a purely photonic state to a collective spin coherence. For read-out, the control field is turned on again and the medium emits a weak pulse, carrying the quantum information contained in the original pulse. While in principle this allows direct mapping of the quantum state of light into long lived coherences in the atomic ground state, the bandwidth of the stored signal is strongly limited by the transparency window associated to EIT.

Various methods have been proposed to achieve broadband memories and escape the limitations linked to EIT. Very interesting protocols are based on the implementation of controlled broadening, such as CRIB (controlled reversible inhomogeneous broadening) [13], using photon echo-type reversal [14] and AFC (atomic frequency comb) where an absorbing comb structure is created in the medium [15]. These techniques have been successfully applied for echo-type light storage in rare-earth doped crystals [16] and atomic vapours [17]. These methods allow broad bandwidth but they imply writing times which are still rather long (of the order of microseconds). In the spatial domain, an interesting phenomena, quantum holography, has been proposed to implement 3D-memories [18,19].

An alternative method is based on the use of a broadband, ultrafast control and signal field pulses for the writing process. It relies on two different atomic transitions sharing the same excited state. The control field and the signal field contribute to a two-photon process coupling two ground states. In this case the signal pulse is converted into an atomic coherence between ground and excited states and then into a ground state coherence by the control pulse. However, since the interaction times are
very short, it does not allow for the buildup of EIT. This method has been proposed [21,22] and demonstrated experimentally in far off-resonance conditions [23,12].

In this paper, we present a detailed theoretical model for an ultrafast memory without adiabatic approximation and valid for arbitrary frequency detuning. We show that this protocol holds the promise for a quantum memory with high efficiency, fast operation and broad bandwidth together with spatial multimode capacity. The problem of the achievable efficiency in this case was treated in Refs. [21,22] using a numerical optimization procedure based on the search of the optimal pulse shape for the signal or driving field in the limit of adiabatic elimination of the excited state. In Ref. [24] the shaping of the driving pulse is based on the analysis of the Lagrange function, avoiding the adiabatic approximation in the optimization procedure. A very good efficiency can be obtained even for short pulse durations. In Ref. [25] a different optimization technique based on the minimization of the losses has been used in the resonant case without adiabatic approximation, in the limit of very short pulses (shorter than the excited state decay time). In the present work we extend this technique to the case of arbitrary frequency detuning.

We show that the memory efficiency can be very good even for large detunings as long as the experimental parameters are properly optimized. Moreover, we explore the transition region between resonant and adiabatic regimes and we demonstrate that our technique allows finding optimal parameters for storage.

The article is organized as follows. In Section 2 we present the model system and we write the main equations ruling it. In Section 3 we give the method for solving the equations for the writing and read-out processes in the semi-classical limit. In Section 4, we study the evolution of the atomic coherence and of the signal field during the writing process. In Section 5, we study the read-out process and the efficiency of the memory as a whole.

2 Model system

In this paper, we consider an ensemble of three-level atoms in a Λ-configuration (Fig. 1) that will be used to store temporal and spatial multimode quantum fields. The atoms interact with two electromagnetic fields, a signal field $E_s$ and a driving field $E_d$, that connect the two atomic ground states to the excited state. The driving field is a strong, classical field propagating as plane wave, while the signal field is a weak quantum field with a transverse structure.

The signal and driving field are very short pulses that are assumed to be much shorter than the excited state lifetime $\gamma^{-1}$, so that we can neglect the spontaneous emission during the writing process.

In the dipole approximation the light matter interaction Hamiltonian is given by

$$\hat{V} = -\sum_j \hat{d}_j(t) \hat{E}(t, \mathbf{r}_j),$$

$$\hat{E}(t, \mathbf{r}_j) = \hat{E}_s(t, \mathbf{r}_j) + \hat{E}_d(t, \mathbf{r}_j).$$  \hspace{1cm} (1)

Here $\hat{d}_j(t)$ is the electric dipole operator of the $j$-th atom located at $\mathbf{r}_j$. In the paraxial and quasi-resonant approximations the Hamiltonian can be rewritten in the form

$$\hat{V} = \ii \int \int \int dz\, d^3\rho \left[ i\hbar g \left( \hat{a}(z, \mathbf{r}, t) \hat{\sigma}_{31}(z, \mathbf{r}, t) e^{ik_z z - i\Delta t} - \hat{a}^\dagger(z, \mathbf{r}, t) \hat{\sigma}_{13}(z, \mathbf{r}, t) e^{-ik_z z + i\Delta t} \right) + \hbar \left( \Omega(t) \hat{\sigma}_{23}(z, \mathbf{r}, t) e^{ik_d z} - i\Delta t \right) - \Omega^*(t) \hat{\sigma}_{23}(z, \mathbf{r}, t) e^{-ik_d z + i\Delta t} \right].$$  \hspace{1cm} (2)

Here $k_z$ and $k_d$ are the wave vectors of the signal and driving fields, and $\mathbf{r} = \mathbf{r}(x, y)$. The one-photon detunings of the signal and driving fields are assumed to be equal, and equal to $\Delta$ so that the two-photon resonance between levels 1 and 2 is fulfilled

$$\Delta = \omega_s - \omega_{13} = \omega_d - \omega_{23}. \hspace{1cm} (3)$$

The spatial coordinates $z$ and $\mathbf{r}$ describe the longitudinal and transverse propagation of the signal field. The normalized amplitude of the signal field $\hat{a}(z, \mathbf{r}, t)$ is written as

$$\hat{E}_s(\mathbf{r}, t) = -i \sqrt{\frac{\hbar \omega_s}{2 \epsilon_0 c}} e^{-i\omega_s t + ik_z z} \hat{a}(z, \mathbf{r}, t) + h.c., \hspace{1cm} (4)$$

where $\hat{a}(z, \mathbf{r}, t)$ is the annihilation operator for the signal field. Under propagation in free space we have the following commutation relations [20]

$$\left[ \hat{a}(z, \mathbf{r}, t), \hat{a}^\dagger(z', \mathbf{r}', t') \right] = \delta^3(\mathbf{r} - \mathbf{r}') \delta(t - t'), \hspace{1cm} (5)$$

$$\left[ \hat{a}(z, \mathbf{r}, t), \hat{a}^\dagger(z', \mathbf{r}', t) \right] =$$

$$= c \left( 1 - i \frac{\partial}{k_z} \frac{\partial}{\partial z} - \frac{c}{2k_z^2} \Delta_{\perp} \right) \delta^3(\mathbf{r} - \mathbf{r}'). \hspace{1cm} (6)$$

The amplitude $\hat{a}(z, \mathbf{r}, t)$ is normalized so that the mean value $\langle \hat{a}^\dagger(z, \mathbf{r}, t) \hat{a}(z, \mathbf{r}, t) \rangle_t$ is the photon number per second per unit area. The symbol $\Delta_{\perp}$ represents the transverse Laplacian with respect to $\mathbf{r}$

$$\Delta_{\perp} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}. \hspace{1cm} (7)$$

![Fig. 1. Three level atomic system interacting with driving field Ω and signal field a.](image-url)
We define the intensity of the driving field through the Rabi frequency $\Omega$. For the sake of simplicity we consider this value as real $\Omega = \Omega^*$. The driving field is a classical plane monochromatic wave propagating along the z-axis. The coupling constant between atom and signal field is

$$g = \left( \frac{\omega_0}{2\epsilon_0hc} \right)^{1/2} d_{31},$$

where $d_{31}$ is the electric dipole element on the transition $|1\rangle \rightarrow |3\rangle$. We define the collective coherences and population as sums of over all the atoms

$$\hat{\sigma}_{ik}(r, t) = \sum_j \hat{\sigma}_{ik}^j(t) \delta^3(r - r_j),$$

$$\hat{N}_i(r, t) = \sum_j \hat{\sigma}_{ik}^j(t) \delta^3(r - r_j).$$

These quantities fulfill the commutation relations

$$[\hat{\sigma}_{ik}(r, t), \hat{\sigma}_{ik}(r', t)] = \left[ \hat{N}_i(r, t) - \hat{N}_k(r, t) \right] \delta^3(r - r').$$

In this basis one can derive a full system of Heisenberg equations for the collective operators namely the field amplitude $\hat{a}(r, t)$, the collective atomic coherences $\hat{\sigma}_{ik}(r, t)$ and the collective atomic populations $\hat{N}_i(r, t)$. Taking into account the Hamiltonian (6) and the commutation relations (9), (11), the complete system of the equations reads

$$\left( \frac{\partial}{\partial t} + c \frac{\partial}{\partial z} - \frac{i c}{2k_s} \Delta_\perp \right) \hat{a} = -cg\hat{\sigma}_{13},$$

$$\frac{\partial}{\partial t} \hat{\sigma}_{13} = -i\Delta \hat{\sigma}_{13} + \Omega \hat{\sigma}_{12} + g\hat{\sigma}(\hat{N}_1 - \hat{N}_3),$$

$$\frac{\partial}{\partial t} \hat{\sigma}_{12} = -\Omega \hat{\sigma}_{13} - g\hat{\sigma}_{12},$$

$$\frac{\partial}{\partial t} \hat{\sigma}_{32} = i\Delta \hat{\sigma}_{32} - \Omega(\hat{N}_3 - \hat{N}_2) + g\hat{\sigma}_{12},$$

$$\frac{\partial}{\partial t} \hat{N}_1 = -g\hat{\sigma}_{31} - g\hat{\sigma}_{13},$$

$$\frac{\partial}{\partial t} \hat{N}_2 = -\Omega(\hat{\sigma}_{32} - \hat{\sigma}_{23}),$$

$$\frac{\partial}{\partial t} \hat{N}_3 = -\frac{\partial}{\partial t} \hat{N}_1 - \frac{\partial}{\partial t} \hat{N}_2.$$

To derive the equations [12]-[18] we have performed the substitutions

$$\hat{\sigma}_{13} \rightarrow e^{-ik з \frac{\partial}{\partial z}} - i\Delta t \hat{\sigma}_{13},$$

$$\hat{\sigma}_{23} \rightarrow e^{-ik з \frac{\partial}{\partial z}} - i\Delta t \hat{\sigma}_{23},$$

$$\hat{\sigma}_{12} \rightarrow e^{i(k у \frac{\partial}{\partial z})} \hat{\sigma}_{12}.$$

We have omitted the terms related to spontaneous relaxation $|3\rangle \rightarrow |1\rangle$ since we have assumed that the relaxation rate $\gamma$ is small enough for the spontaneous emission to be negligible during the short time duration of the pulses. The equations can be written in a simplified way with a few approximations given below.

According to equations [10] the collective atomic operators have sharp spatial distributions due to the delta-localization of the atoms. However, due to the collective effect of atoms localized along a field trajectory, we can average Eqs. [12]-[18] over the positions of the atoms.

We will also replace the operator $\hat{N}_1 - \hat{N}_3$ in Eq. (13) by the number giving the mean atomic density $N$. We will assume that in the beginning of the process most atoms are in state $|1\rangle$. During the memory processes (writing and read-out) the population of the state $|1\rangle$ stays close to its initial value, because the photon number in the signal pulse is much smaller than the initial atomic number.

In Eq. (14), we can neglect the second term on the right hand side since $g\hat{\sigma}$ is much smaller than $\Omega$ (we have assumed $|\Omega|^2 \gg g^2(\hat{a}^\dagger \hat{a})$). Furthermore, $\hat{\sigma}_{32} \ll \hat{\sigma}_{13}$ because the populations $N_2$ and $N_3$ are much smaller than $N$.

Now we can write a simplified system of partial differential equations describing the evolution of the system as

$$\left( \frac{\partial}{\partial t} + c \frac{\partial}{\partial z} - \frac{i c}{2k_s} \Delta_\perp \right) \hat{a}(z, \rho, t) = -g \hat{\sigma}_{13}(z, \rho, t),$$

$$\frac{\partial}{\partial t} \hat{\sigma}_{13}(z, \rho, t) = -i\Delta \hat{\sigma}_{13}(z, \rho, t) + g\hat{\sigma}(\hat{N}_1 - \hat{N}_3),$$

$$\frac{\partial}{\partial t} \hat{\sigma}_{12}(z, \rho, t) = -\Omega \hat{\sigma}_{13}(z, \rho, t) - g\hat{\sigma}_{12},$$

$$\frac{\partial}{\partial t} \hat{\sigma}_{32}(z, \rho, t) = i\Delta \hat{\sigma}_{32}(z, \rho, t) - \Omega(\hat{N}_3 - \hat{N}_2) + g\hat{\sigma}_{12},$$

$$\frac{\partial}{\partial t} \hat{\sigma}_{13}(z, \rho, t) = -\frac{\partial}{\partial t} \hat{\sigma}_{12} - \frac{\partial}{\partial t} \hat{\sigma}_{32},$$

$$\frac{\partial}{\partial t} \hat{N}_1 = -g\hat{\sigma}_{31} - g\hat{\sigma}_{13},$$

$$\frac{\partial}{\partial t} \hat{N}_2 = -\Omega(\hat{\sigma}_{32} - \hat{\sigma}_{23}),$$

$$\frac{\partial}{\partial t} \hat{N}_3 = -\frac{\partial}{\partial t} \hat{N}_1 - \frac{\partial}{\partial t} \hat{N}_2.$$

In a previous publication [25], similar equations were derived for the case of a resonant excitation ($\Delta = 0$). Here and below we omit the averaging over the atomic localizations.

Let us renormalize the coherences $\hat{\sigma}_{13}$ and $\hat{\sigma}_{12}$

$$\hat{\sigma}_{12}(z, \rho, t)/\sqrt{N} = \hat{b}(z, \rho, t),$$

$$\hat{\sigma}_{13}(z, \rho, t)/\sqrt{N} = \hat{c}(z, \rho, t)$$

so that they obey to the bosonic commutation relations:

$$[\hat{b}(r, t), \hat{b}^\dagger(r', t)] = [\hat{c}(r, t), \hat{c}^\dagger(r', t)] = \delta^3(r - r').$$

Here again we have taken into account the fact that $\hat{N}_1 - \hat{N}_{2,3} \rightarrow N$.

In the Fourier domain relative to the transverse coordinates $\rho$ the equations read

$$\frac{\partial}{\partial z} \hat{a}(z, t; q) = -g\sqrt{N} \hat{c}(z, t; q),$$

$$\frac{\partial}{\partial t} \hat{c}(z, t; q) = -i\Delta \hat{c}(z, t; q)$$

$$+ g\sqrt{N} \hat{a}(z, t; q) + \Omega \hat{b}(z, t; q),$$

$$\frac{\partial}{\partial t} \hat{b}(z, t; q) = -\Omega \hat{c}(z, t; q).$$
where we have introduced the transverse wavenumber \( q \) and we have made the changes
\[
\hat{a}(z, t; q) \rightarrow \hat{a}(z, t; q) e^{-i q^2 z / (2 k_s)}, \quad (29)
\]
\[
\hat{b}(z, t; q) \rightarrow \hat{b}(z, t; q) e^{-i q^2 z / (2 k_s)}, \quad (30)
\]
\[
\hat{c}(z, t; q) \rightarrow \hat{c}(z, t; q) e^{-i q^2 z / (2 k_s)}. \quad (31)
\]
From the system (26)-(28) one can obtain a conservation equation
\[
\frac{\partial \hat{a}^\dagger}{\partial z} + \frac{\partial \hat{b}^\dagger}{\partial t} + \frac{\partial \hat{c}^\dagger}{\partial t} = 0. \quad (32)
\]
This equation means that the input photons of the weak quantum field are converted into excitations of the atomic coherence \( \hat{\sigma}_a \) and \( \hat{\sigma}_b \). The aim is to store the information carried by the signal field in the ground state coherence \( \hat{\sigma}_b \), so that an excitation of the state \( |3\rangle \) is undetected. It is possible to reduce this loss channel by an appropriate choice of the driving field. If the driving field power is high enough for the Rabi oscillation on the transition \( |2\rangle \rightarrow |3\rangle \) to be more effective than a spontaneous emission (\( \Omega \gg \gamma \)) and if the pulse duration is short enough, then the atoms undergoing a Rabi oscillation have no time to go back to the state \( |3\rangle \), then the third term in Eq. (32) should be negligible. This will be studied in the optimization of the memory process.

We neglect the time delay linked to the pulse propagation in the atomic medium. This means that, if we have long enough pulses, such that \( L/c \ll T \) (\( L \) is the thickness of the medium and \( T \) is the pulse duration), we can neglect the time interval between the time at which the front part of the pulse enters the medium and the time at which the front part leaves it. Formally this means we can neglect the time derivative in Eq. (20). For simplicity we will assume that the driving pulse has a rectangular time distribution (in the equations \( \Omega(t) = \text{const} \) for \( 0 < t < T \)).

### 3 Writing and read-out processes in the semi-classical limit

The main aim of this paper is the determination of the memory efficiency. As it is well known the semiclassical description is sufficient for this and we can use the initial conditions : \( b(0, z; q) = c(0, z; q) = 0 \) for writing and \( a_{in}(t; q) = c(0, z; q) = 0 \) for read-out. Here and below we omit the operator notation for the variables. The detailed resolution of the system of partial differential equations (26)-(28) can be found in Apps. A,B,C.

Using the general solutions (51)-(53) one can obtain the semi-classical ones for the writing process for both the field amplitude \( a^W(t, z; q) \) and the atomic coherence \( b(t, z; q) \) in the form
\[
a^W(\tilde{t}, \tilde{z}; q) = \int_0^{\tilde{T}_W} d\tilde{t} a_{in}(\tilde{t}, q) G_{aa}(\tilde{t} - \tilde{t}', \tilde{z}), \quad (33)
\]
\[
b^W(\tilde{t}, \tilde{z}; q) = -p \int_0^{\tilde{T}_W} d\tilde{t} a_{in}(\tilde{t}, q) G_{ab}(\tilde{t} - \tilde{t}', \tilde{z}), \quad (34)
\]
where we have introduced the dimensionless time \( \tilde{t} \) and longitudinal spatial coordinate \( \tilde{z} \) according to
\[
\tilde{t} = \frac{\Omega t}{L}, \quad \tilde{T}_W = \frac{\Omega T_W}{L}, \quad \tilde{z} = \frac{2 q^2 N}{\Omega} \tilde{z}, \quad \tilde{L} = \frac{2 q^2 N}{\Omega} \tilde{L} \quad (35)
\]
where \( T_W \), is the common duration of signal and driving pulses for writing and \( L \) is the thickness of the memory cell. We have defined an effective interaction coefficient \( p \) given by
\[
p = \frac{g \sqrt{N}}{\Omega}. \quad (36)
\]

The kernels \( G_{aa}(t, z) \) and \( G_{ab}(t, z) \) are time convolutions
\[
G_{aa}(\tilde{t}, \tilde{z}) = \int_0^{\tilde{t}} d\tilde{t}' f(\tilde{t}', \tilde{z}; r) f^*(\tilde{t}' - \tilde{t}, \tilde{z}; -r), \quad (37)
\]
\[
G_{ab}(\tilde{t}, \tilde{z}) = \int_0^{\tilde{t}} d\tilde{t}' f_0(\tilde{t}', \tilde{z}; r) f_0^*(\tilde{t}' - \tilde{t}, \tilde{z}; -r), \quad (38)
\]
where the functions \( f \) and \( f_0 \) are expressed via the n-th Bessel function of the first kind denoted by \( J_n \) :
\[
f(\tilde{t}, \tilde{z}; r) = \delta(\tilde{t}) e^{-ir\tilde{t}} e^{-i \sqrt{1 + r^2} \tilde{t}} \times \sqrt{\frac{(1 + r^2)\tilde{z}}{4\tilde{t}}} J_1 \left( \sqrt{(1 + r)\tilde{z}\tilde{t}} \right) \Theta(\tilde{t}), \quad (39)
\]
\[
f_0(\tilde{t}, \tilde{z}; r) = e^{-ir\tilde{t}} e^{-i \sqrt{1 + r^2} \tilde{t}} J_0 \left( \sqrt{(1 + r)\tilde{z}\tilde{t}} \right) \Theta(\tilde{t}). \quad (40)
\]
Here the frequency detuning is given by the dimensionless parameter \( r = \Delta / (2\Omega) \) and \( \Theta(\tilde{t}) \) is the step function \( \Theta(\tilde{t}) = 1 \) for \( 0 < \tilde{t} < \tilde{T}_W \) and equals zero otherwise.

In order to estimate the efficiency of the memory, we calculate the field amplitude at the output of the cell. This amplitude is different in the case of forward and backward retrieval. The corresponding expressions read
\[
a_{fw}(\tilde{t}, \tilde{L}; q) = \quad (41)
\]
\[
= \frac{1}{2} \int_0^{\tilde{T}_W} d\tilde{t} a_{in}(\tilde{t}, q) \int_0^{\tilde{L}} d\tilde{z} G_{ab}(\tilde{t}, \tilde{z}) G_{ba}(\tilde{t}, \tilde{L} - \tilde{z}), 
\]
\[
G_{ba}(\tilde{t}, \tilde{z}) = G_{ab}(\tilde{t}, \tilde{z}), \quad (35)
\]
and
\[
a_{bw}(\tilde{t}, \tilde{L}; q) = \quad (42)
\]
\[
= \frac{1}{2} \int_0^{\tilde{T}_W} d\tilde{t} a_{in}(\tilde{t}, q) \int_0^{\tilde{L}} d\tilde{z} G_{ab}(\tilde{t}, \tilde{z}) G_{ba}(\tilde{t}, \tilde{L}).
\]

The last two formulas are correct only in the approximation where diffraction is neglected. Indeed, we have not taken into consideration the diffraction phenomenon described by equation (29)-(31). As was demonstrated in (25) this effect does not introduce any significant corrections in the case of the forward retrieval but restricts the mode number for the backward retrieval.
In the following we will use an optimization procedure based on the choice of the optimal relation between the thickness of the memory cell $L$ and the duration $T_W$ of the signal and driving pulses. We will compare this optimization approach with the approach developed in Ref. 22 based on the optimization of the driving pulse shape in the case of short pulses.

4 Discussion of the writing process

In this section, we study the writing process, i.e. the conversion of the signal field into atomic coherence. Let us start with a simple calculation of this process at the input of the memory cell. For this we solve Eq. (44) at $\tilde{z} = 0$. In this case the atomic polarization $b^{W}(\tilde{t}, 0; \tilde{q})$ corresponding to the "written" information reads

$$b^{W}(\tilde{t}, 0) = b^{W}(\tilde{t}, 0; \tilde{q}) / (-2pa_{n}(\tilde{q})) =$$

$$= \frac{1}{2} \left( 1 - e^{-i\tilde{q}} \cos(\sqrt{\tilde{t}^2 + 1}) + \frac{i\tilde{r}}{\sqrt{\tilde{t}^2 + 1}} \sin(\sqrt{\tilde{t}^2 + 1}) \right).$$

For the sake of the simplicity we have taken $a_{n}(\tilde{t}, \tilde{q}) = \text{const}(\tilde{t})$. The normalized atomic polarization ranges from 0 to 1. In the limit of small or large detuning, we have a simple behaviour

$$r \ll 1 : \quad \left| b^{W}(\tilde{t}, 0) \right|^2 = \sin^4 \frac{\tilde{t}}{2},$$

$$r \gg 1 : \quad \left| b^{W}(\tilde{t}, 0) \right|^2 = \sin^2 \frac{\tilde{t}}{4r}.$$

Fig. 2a shows the Rabi oscillation for a normalized detuning $r = 0$ while Fig. 2b shows a slow oscillation (that corresponds to the term with frequency $\sqrt{\tilde{t}^2 + 1} - r$ in (43)), modulated by a fast oscillation (at frequency $\sqrt{\tilde{t}^2 + 1} + r$), with a modulation depth decreasing with increasing $r$.

Following the variation of $\left| \tilde{b}^{W}(\tilde{t}, 0) \right|^2$ as a function of $r$ allows to get a first view of the behaviour of the system when the detuning is varied. It can be seen from Fig. 2a that even for a small detuning, $r = 0.1$, a significant distortion of the coherence profile at the input of the medium takes place as compared to the resonant case. On the other hand for $r = 2$ the high frequency modulation of the slow oscillations is rather small and the excitation can be considered as close to the off-resonant case, where the well-known solutions in the Raman limit can be used. A detailed study of the coherence distribution for all $\tilde{z}$, as given below, will allow to better characterize the interaction regime, between resonant and Raman.

When shifting into the medium, the behavior of the coherence is much more complicated since it loses its simple harmonic character. Figure 3 shows a displacement of the maximum of the coherence along the medium. Let us follow the dependence of $\left| \tilde{b}^{W}(\tilde{t}, \tilde{z}) \right|^2$ on $\tilde{z}$ for a given value of $\tilde{t} = \tilde{T}_W = \pi$ given in Fig. 3.

Comparing the curves for different values of the detuning, we see that the value of coherence at $\tilde{z} = 0$ significantly decreases with increasing detuning. From these curves it could be concluded that the larger $r$ the less information is written in the cell. However, we will show that it is not true. Let us introduce the quantity $n^{\epsilon IF}$ that characterizes the proportion of signal photons converted into coherence $\tilde{b}^{W}$ during the writing process:

$$n^{\epsilon IF}(\tilde{T}_W, \tilde{L}) = \frac{1}{\tilde{T}_W} \frac{1}{2} \int_{0}^{\tilde{L}} \left| \tilde{b}^{W}(\tilde{T}_W, \tilde{z}) \right|^2 d\tilde{z}.$$ (46)

Here, $1/\tilde{T}_W$ before the integral comes from the input pulse energy, and the factor $1/2$ comes from the previously introduced dimensionless variables. The integral gives the normalized population $N_2$ in the medium with length $L$ during the interaction time $T_W$. Since the transition of an atom to the level $|2>$ in our model corresponds to the coherent scattering of a photon from the signal wave, this is also the number of signal photons that was recorded in the atomic coherence. We are interested in the percentage of input photons recorded in such a way. The calculated values of $n^{\epsilon IF}$ are presented Fig. 3c. We see that the number of recorded photons in the first two panels are almost identical (there is even a small increase of $n^{\epsilon IF}$ for $r = 0.5$ as compared with the resonant value), while for $r = 2$ the value $n^{\epsilon IF}$ decreases by a factor of less than 2.

We can also follow the dependence of $n^{\epsilon IF}$ on $r$ for a given medium length and for different durations of the signal pulse. Fig. 3 shows that the proportion of recorded photons decreases when the detuning increases, but this decrease depends on the duration of the writing process. In particular, we see that the writing efficiency depends weakly on the detuning for $\tilde{T}_W = \pi/4$ but it stays quite low. With increasing pulse duration ($\tilde{T}_W = \pi$) a plateau appears for detuning range 0 to 0.7. This range is reduced for $\tilde{T}_W = 2\pi$, but a second plateau appears for $r \in [1,1.8]$. While Figs. 3a and 3b give a detailed behaviour of the efficiency of the writing process depending on the detuning, on the time duration of the pulse and on the length of the medium, it appears that the optimization of the efficiency is non trivial and requires a specific procedure. This will be studied in the next section.

4.1 Estimation of the writing losses

In Ref. 23 an algorithm of memory optimization, based on the minimization of leakage is described. Leakage is defined as

$$L(\tilde{T}_W, \tilde{L}) = \frac{1}{\tilde{T}_W} \int_{0}^{\tilde{T}_W} |a_{\tilde{n}}(\tilde{t})|^2 d\tilde{t} \times 100\%.$$ (47)

It characterizes the proportion of signal photons going out of the cell during the writing time. Such an estimation of
the medium. However, as will be shown below, there is a range of $T^W$ in which the population of the upper energy level is large enough to cause significant losses. We will show that for a high-speed memory the three-level atomic system can not be reduced to a two-level scheme. Note that this situation is specific of the case of instantaneous interaction of the signal and control fields with matter, and does not happen in memory protocols based on an echo [15]. We have already introduced the value $\eta_{eff}$, which is the proportion of signal photons that have been recorded. Then, the value of the total losses (as a percentage of the number of photons in the input signal pulse) can be expressed as follows:

$$\mathcal{L}(\tilde{T}^W, \tilde{L}) = (1 - \eta_{eff}(\tilde{T}^W, \tilde{L})) \cdot 100\%.$$  \hspace{1cm} (48)

Fig. 6 shows the losses associated with leakage $\mathcal{L}$ (blue curves, dotted lines) and the total losses of photons $\mathcal{L}_c$, (red curves, full lines) as a function of the duration of the writing for a given medium length, and three different values of $r$.

First, let us note that these curves are not monotonous and exhibit one or several minima. This means that for a given length $\tilde{L}$ one can find the pulse duration which is recorded in the atomic medium with minimal losses. Significant difference between the curves corresponding to leakage and to total losses come from the role of the upper level in the interaction of such pulses with the atomic medium. However, in the region of minimum losses, the distance between the two curves is small. The optimization of the memory based on leakage allows to define a range of values $T^W$, for which an efficient writing is expected. However, the curve giving the total losses is a more precise tool to determine the optimum ratio between $T^W$ and $\tilde{L}$. As can be seen from the plots, above some value of $T^W$ the two curves coincide, i.e. all the system losses are associated with leakage only and level of $|3\rangle$ is not populated at the end of the process. The larger detuning $r$ the smaller the value $T^W$ for which this happens.

We can also follow the dependence of the losses on the length $\tilde{L}$ for a given value of $T^W$ as shown in Fig. 7. These curves have a monotonous variation, and show that the efficiency increases with increasing medium length. When the pulse duration increases, the curves of $\mathcal{L}$ and $\mathcal{L}_c$ get closer to each other, in agreement with the analysis of Fig. 6. A specific feature of these curves is their saturation for large values of $\tilde{L}$. The presence of a plateau on the leakage plots is actually due to an approximation made in the model; we have neglected the time intervals associated with the propagation of the pulse wavefronts inside the medium, and we assumed $t = 0$ is the time at which the wavefront reaches the cell output, while $\tilde{t} = \tilde{T}^W$ is the time at which the end part of the pulse arrives at the entrance of the cell. This means that we always have some leakage in the initial time independently of the length of the medium.

The difference between the levels of the plateau for $\mathcal{L}$ and $\mathcal{L}_c$ characterizes the losses due to the non zero population of level $|3\rangle$. One can see that this value is constant for large enough length $\tilde{L}$. This value saturates because of the depletion of the signal field, so that further increase of the medium length cannot change the populations $N_2$ and $N_3$. The difference between $\mathcal{L}$ and $\mathcal{L}_c$, depends strongly on the pulse duration. In particular if the pulse is too short, many atoms are left in the upper state. It can also be seen in Figs. 7 that when the detuning increases, the saturation occurs at larger values $\tilde{L}$, and that a high efficiency can be reached as well if the atomic medium is long enough. Various optimizations procedures have been proposed in the limit when the excited state can be adiabatically eliminated [21,22]. In particular, the optimization used in reference [22] allows to get the maximum available efficiency for long enough durations of the writing process, but breaks down when the duration of a writing process $T^W$ gets smaller than the excited state decay time $\gamma^{-1}$ divided by the optical depth of the medium $d$.

In Ref. [24] the numerical optimization procedure relying on the shaping of the driving pulse was extended to the non adiabatic case, which allowed to reach better storage efficiency for short pulses. The latter technique was developed for the resonant case, yielding optimal memory efficiencies that are very close to the ones presented here. The applicability of our optimization method to various detunings shows that such a memory can be also very efficient in the off-resonant regime, bringing more flexibility for experimental realizations.

4.2 Validity limits of resonant and Raman approximations

The solutions that we have presented are valid for detunings ranging from zero to large values that correspond to the case of Raman interaction, where the system is effectively reduced to a two-level system. General formulas covering the full range of detunings allow a comparison with the limit cases of resonant and Raman interactions. We can identify the largest detuning for which the resonant approximation is still valid, yielding the same storage efficiency. On the other hand, we can estimate for which value of $\Delta$ a simplified Raman description can be used without yielding appreciable errors in the memory efficiency.

In Fig. 8a, one can see a significant distortion of the temporal profile of the atomic ground state coherence at the input of the medium for $r = 0.1$ as compared to the resonant case. However this detuning does not affect the writing efficiency in a significant way. Figure 8b shows the dependence of the total losses on time for a given length in two cases: for exact resonance (blue curve, dotted line) and with a detuning $r = 0.1$ (red curve, full line). The curves coincide to within 1.5% over the full range of $\tilde{T}^W$. Thus, despite the local differences in the field-atom interaction in these two cases, the presence of a small detuning does not actually change the properties of memory cell as a whole. However when the detuning increases, the difference between the curves increases (at $r = 0.2$ it reaches 4.5%, see Fig. 8b), but in the range of interest for $\tilde{T}^W$, that is the one that allows minimization of the losses, the curves are still close to each other (they agree within 1.5%).
increase of the detuning distorts the profile of the losses further, and the value of $T^W$ that provides minimum losses is shifted (see Fig. 5a-d).

Let us now turn to the case of large detuning. One can see in Fig. 9 that for $r = 3$ the profile of the total losses calculated with the general formulas (33)-(34) coincides well with the profile for the same quantity, calculated in the Raman approximation with $r \gg 1$ (the difference is about 2.5%, and it is less than 1% at the minimum). For a smaller detuning (for $r = 2$) the difference between the curves increases up to 7% (about 3% at the minimum): the calculation made in the Raman case underestimates the minimum losses and shifts toward lower values of $T^W$. Thus, we can conclude that for $r = 3$ and higher the Raman approach is applicable with good accuracy, but the general solutions should be used for lower values of the detuning.

For large enough detunings where the adiabatic limit is valid as well as our model, we can compare the results further. Our model predicts a storage efficiency below the maximal available efficiency reached by the method proposed in reference [22]. In the third plot of Fig. 6 the efficiency is 65% for $Td\gamma = T^W L/2 = 30$ and $d = 2400$. For the same parameters the adiabatic optimization method converges to the maximal efficiency which is close to 100%. This is due to the fact that we do not elaborate shaping of a control pulse, used in reference [22]. Thus our method, even without control pulse shaping, is quite powerful in a non-adiabatic limit. Otherwise, the numerical adiabatic optimization of the control pulse profile should be used to reach the maximal storage efficiency for the long pulses.

For this value of $T^W$ we calculate the intensity of the retrieval field (normalized to the intensity of the input signal) as a function of the reading time (see Fig. 10b). The plot obtained from optimization based on leakage [25] is shown in Fig. 10a, in order to compare the results. The retrieval efficiency is defined by:

$$E(q) = \frac{\int_0^{T^R} |a_R(\tilde{t}, \tilde{q})|^2 d\tilde{t}}{\int_0^{T^W} |a_{in}(\tilde{t}, \tilde{q})|^2 d\tilde{t}} \times 100\%$$  (49)

In the case of optimization based on total losses, we find that the retrieval efficiency is equal to 88% at $T^R = 2T^W$. In view of the writing efficiency ($\mathcal{L}_c = 11.2\%$), we see that using a reading time $T^R = 2T^W$ we can restore almost all the information written in the medium. When the leakage-based optimization is used, we find $E = 84\%$ at $T^R = 3T^W$. This comparison clearly demonstrates the benefits of the optimization based on the total losses.

As for the temporal profiles of retrieval field, it is obvious that in both cases they are very different from the input signal profile, which is a usual result in most memory processes.

Let us now consider the result of the forward and backward retrieval for the non-resonant case (see Fig. 11 for $r = 0.5$).

The efficiency is much lower for forward retrieval than for backward retrieval one: at $T^R = 10$ (i.e. $T^R \approx 3.3T^W$) the efficiency of forward process is $E = 58\%$, while the efficiency of backward process is $E = 85.6\%$. Moreover, in the latter case less than 2% of the available photons remain in atomic ensemble (since the writing losses were 12.6%). Comparison with the resonant case shows that the presence of detuning slows down the read-out.

Finally, we can show that even for a large detuning, the total efficiency can be large, at the condition that the medium length $L$ is large enough. In Fig. 12 the red (oscillating) curve corresponds to the calculated read-out signal intensity for backward retrieval for $L = 100$, $r = 2$, and for a pulse duration of the input field $T^W = 4\pi$, which ensures a minimum in the total writing losses (writing efficiency is $v_{eff} = 95.6\%$). The oscillation period is equal to $4\pi$, similar to the modulation observed in Fig. 2. If we calculate the intensity of the retrieved signal in the Raman approximation, the overall shape of the curve remains the same, but the oscillations disappear (blue curve, dotted line in Fig. 12). The retrieval efficiency calculated with these curves differs by 1.4% (78.6% for the exact calculation and 77.2% for the calculation in Raman approximation), so that the calculation in Raman limit can be a quite good estimation for memory efficiency at $r = 2$. Note, that here like for the curves in Fig. 2 the magnitude of the oscillations will decrease with increasing detuning.

In addition to high speed operation, our model includes transverse coordinates, as can be seen from Eq. (2), and thus allows the treatment of a variety of multimode fields.

5 Discussion of the read-out process

In Ref. [25] the optimization of the retrieval efficiency was based on the choice of the pulse duration providing the minimum leakage for a given medium length. Here we will look for a minimization of the total losses. Moreover, as was shown in Ref. [22][25], backward retrieval provides significantly larger efficiency than forward retrieval, therefore we will study the cases of forward and backward retrieval and compare them. Using the result of Fig. 6, we choose a value for input signal duration of $T^W = 5.5$ that provide minimum total losses for the writing process.

For this value of $T^W$ we calculate the intensity of the retrieval field (normalized to the intensity of the input signal) as a function of the reading time (see Fig. 10b). The plot obtained from optimization based on leakage [25] is shown in Fig. 10a, in order to compare the results. The retrieval efficiency is defined by:

$$E(q) = \frac{\int_0^{T^R} |a_R(\tilde{t}, \tilde{q})|^2 d\tilde{t}}{\int_0^{T^W} |a_{in}(\tilde{t}, \tilde{q})|^2 d\tilde{t}} \times 100\%$$  (49)

In the case of optimization based on total losses, we find that the retrieval efficiency is equal to 88% at $T^R = 2T^W$. In view of the writing efficiency ($\mathcal{L}_c = 11.2\%$), we see that using a reading time $T^R = 2T^W$ we can restore almost all the information written in the medium. When the leakage-based optimization is used, we find $E = 84\%$ at $T^R = 3T^W$. This comparison clearly demonstrates the benefits of the optimization based on the total losses.

As for the temporal profiles of retrieval field, it is obvious that in both cases they are very different from the input signal profile, which is a usual result in most memory processes.
in the same way as in Ref. [25]. The storage of a multimode signal field is performed using a single mode signal field, and the information is written in the atomic ensemble as a hologram. The quantum hologram process together with a Raman memory was proposed for the first time by Sokolov et al. in Ref. [20] and it was shown that the memory capacity is limited by diffraction, but with the limitation depending on the direction of readout. Under forward readout the maximum number of the modes or pickels $N$ which can be stored is in principle given by the square of the Fresnel number $N \sim F_N^2$, where $F_N = S/(\lambda L)$. This expresses the condition that the output pixel size should not exceed the transverse size of the memory cell. This rather loose limitation comes from the fact that most diffraction effects are compensated between writing and readout with similar geometries. A similar compensation does not take place for the backward readout and as a result the number of stored modes $N$ in this case cannot exceed the Fresnel number.

6 Conclusion

In this article, we have presented a model for a high speed quantum memory based on a three-level medium in the $\Lambda$-configuration. This model relies on a full calculation of the fields and of the atomic coherences as a function of time and space for arbitrary frequency one-photon detuning, while keeping the two-photon resonance. It allows to examine all the situations between interactions resonant with the excited state and non resonant interaction, which corresponds to a Raman transition. Our model allows to identify the conditions in which the interaction can be treated as a Raman transition. This corresponds to a normalized detuning $r = \Delta/(2\Omega) > 2$. For larger detunings, we have shown that the Raman model gives accurate results.

In the near-resonant case, the interaction of the three-level atomic system with the weak signal field and the strong driving field turns out to be more complicated than in the Raman case. As a matter of fact in contrast to the Raman process, where the upper atomic state is practically not involved in the process, all three levels are populated, and this leads ultimately to additional undesirable losses. We have shown that it is possible to choose the parameters to make these losses very small. By controlling the Rabi oscillation one can preferentially populate the lower state $|2\rangle$ and depopulate state $|3\rangle$. When the frequency detuning increases, the upper state population naturally decreases, which contributes to the reduction of losses. Moreover, for large frequency detunings, the interaction between the atomic medium and the fields is weaker, and a larger medium depth is necessary to reduce the leakage.

An important question is how to optimize the memory process to obtain the highest possible quantum efficiency. In the literature two approaches are mainly discussed. The first one is based on the choice of the optimal time-shape for the signal pulse, using for this the eigenfunctions of the integral operators specific of the considered memory [21]. It was demonstrated that if the functions possess some symmetry, the efficiency can be improved. In our case with very short pulses, there is no such symmetry and this optimization turns out to be impossible. Another optimization method was proposed by Gorshkov et al. in Ref. [22], based on a study of the driving pulse shape. However, the search procedure for the optimal shape in this article is based on the adiabatic or Raman approximations, which are not generally applicable in the case of very short pulses. Non-adiabatic pulse shape optimization was also developed for the resonant case [23], extending the validity of the method. In our model, the optimization is based on the minimization of the losses, which come from the leakage and from the population of the upper atomic level, and allows identifying the optimal combination of signal pulse duration and optical depth. We demonstrate that high memory efficiencies can be achieved by this method for very short pulses whatever the value of the detuning.

7 Acknowledgements

The study was performed within the framework of the Russian-French Cooperation Program "Lasers and Advanced Optical Information Technologies", of the European Project HIDEAS (grant No. 221906). O.M. acknowledges the support of Ile de France programme IFRAF. The study was also supported by RFBR (grant No. 08-02-92504).

A General solutions for the main equations

Eqs. (26)-(28) can be solved in the general form by using the Laplace domain. Some details of the formal procedures are discussed in App. [4] and now we start with the solutions themselves in the explicit form. For our analysis in this article we do not need full information about the solutions, nevertheless one can find it below. Here and everywhere in article under consideration of the solution we use dimensionless co-ordinates $\bar{t}$ and $\bar{z}$ given by

$$\bar{t} = \Omega t, \quad \bar{z} = \frac{2\gamma^2 N}{\Omega} z. \quad (50)$$
Let us write the general solutions in the form under the arbitrary initial and boundary conditions
\[
\hat{a}(\tilde{t}, \tilde{z}; q) = \int_0^\infty d\tilde{\tau} \tilde{a}_{in}(\tilde{t} - \tilde{\tau}; q)G_{ao}(\tilde{t}, \tilde{z}) - \frac{1}{2p} \int_0^\infty d\tilde{\tau} \hat{b}(0, \tilde{z} - \tilde{\tau}; q)G_{ba}(\tilde{t}, \tilde{z}') - \frac{1}{2p} \int_0^\infty d\tilde{\tau} \hat{c}(0, \tilde{z} - \tilde{\tau}; q)G_{ca}(\tilde{t}, \tilde{z}'), \tag{51}
\]
\[
\hat{b}(\tilde{t}, \tilde{z}; q) = -p \int_0^\infty d\tilde{\tau} \tilde{a}_{in}(\tilde{t} - \tilde{\tau}; q)G_{ab}(\tilde{t}', \tilde{z}) + \frac{1}{2p} \int_0^\infty d\tilde{\tau} \hat{b}(0, \tilde{z} - \tilde{\tau}; q)G_{bb}(\tilde{t}, \tilde{z}') + \frac{1}{2p} \int_0^\infty d\tilde{\tau} \hat{c}(0, \tilde{z} - \tilde{\tau}; q)G_{bc}(\tilde{t}, \tilde{z}'), \tag{52}
\]
\[
\hat{c}(\tilde{t}, \tilde{z}; q) = p \int_0^\infty d\tilde{\tau} \tilde{a}_{in}(\tilde{t} - \tilde{\tau}; q)G_{ac}(\tilde{t}', \tilde{z}) + \frac{1}{2p} \int_0^\infty d\tilde{\tau} \hat{b}(0, \tilde{z} - \tilde{\tau}; q)G_{bc}(\tilde{t}, \tilde{z}'), \tag{53}
\]
where kernels \(G_{ak}(\tilde{t}, \tilde{z})\) are bilinear combinations of the expressions depending on the \(n\)-th Bessel functions of the first kind denoted by \(J_n\)

\[
f(\tilde{t}, r; \tilde{r}) = \delta(\tilde{t}) - e^{-i \left(\frac{1+\tilde{t}^2}{2}\right) \tilde{t}} \left(\frac{1}{\sqrt{1+\tilde{t}^2}}\right) J_0(4\tilde{t}) J_1(\sqrt{2+1} \tilde{t}, \Theta(\tilde{t})), \tag{54}
\]
\[
f_1(\tilde{t}, r; \tilde{r}) = e^{-i \left(\frac{1+\tilde{t}^2}{2}\right) \tilde{t}} \left(\frac{1}{\sqrt{1+\tilde{t}^2}}\right) J_0(4\tilde{t}) J_1(\sqrt{2+1} \tilde{t}, \Theta(\tilde{t})), \tag{55}
\]
\[
f_0(\tilde{t}, r; \tilde{r}) = e^{-i \left(\frac{1+\tilde{t}^2}{2}\right) \tilde{t}} \left(\frac{1}{\sqrt{1+\tilde{t}^2}}\right) J_0(4\tilde{t}) J_1(\sqrt{2+1} \tilde{t}, \Theta(\tilde{t})). \tag{56}
\]

The kernels in Eq. \((51)\) read
\[
G_{aa}(\tilde{t}, \tilde{z}) = [f(r) * f^*(−r)](\tilde{t}, \tilde{z}), \tag{57}
\]
\[
G_{ba}(\tilde{t}, \tilde{z}) = [f_0(r) * f_0^*(−r)](\tilde{t}, \tilde{z}), \tag{58}
\]
\[
G_{ca}(\tilde{t}, \tilde{z}) = \frac{1}{2} [f(r) * f^*(−r)](\tilde{t}, \tilde{z}) + \frac{1}{2} [f_0(r) * f_0^*(−r)](\tilde{t}, \tilde{z}). \tag{59}
\]

We denote the \(t\) convolution of two arbitrary functions \(X(\tilde{t}, \tilde{z}; r)\) and \(Y(\tilde{t}, \tilde{z}; r)\) as
\[
[X(r) * Y^*(−r)](\tilde{t}, \tilde{z}) = \int_0^\infty d\tilde{\tau} X(\tilde{t} - \tilde{\tau}, \tilde{z}; r) Y^*(\tilde{\tau}, \tilde{z}; −r). \tag{60}
\]

The parameter \(r\) is the dimensionless frequency detuning \(r = \Delta/(2\Omega)\).

The other kernels read
\[
G_{ab}(\tilde{t}, \tilde{z}) = [f_0(r) * f_0^*(−r)](\tilde{t}, \tilde{z}), \tag{61}
\]
\[
G_{bc}(\tilde{t}, \tilde{z}) = 2 \delta(\tilde{z}) F_1(\tilde{t}) + [f_1(r) * f_1^*(−r)](\tilde{t}, \tilde{z}), \tag{62}
\]
\[
G_{ab}(\tilde{t}, \tilde{z}) = 2 \delta(\tilde{z}) F_2(\tilde{t}) + \frac{1}{2} [f_1(r) * f_1^*(−r)](\tilde{t}, \tilde{z}) \tag{63}
\]
\[
G_{bc}(\tilde{t}, \tilde{z}) = 2 \delta(\tilde{z}) F_3(\tilde{t}) - \frac{1}{2} [f_1(r) * f_1^*(−r)](\tilde{t}, \tilde{z}), \tag{64}
\]
\[
G_{bc}(\tilde{t}, \tilde{z}) = \frac{1}{2} [f_0(r) * f_0^*(−r)](\tilde{t}, \tilde{z}), \tag{65}
\]
\[
G_{bc}(\tilde{t}, \tilde{z}) = \frac{1}{4} [f_1(r) * f_1^*(−r)](\tilde{t}, \tilde{z}) \tag{66}
\]

In these formulas the time dependent factors \(F(\tilde{t})\) read
\[
F_1(\tilde{t}) = \left[\cos\left(\frac{1+\tilde{t}^2}{2}\right) \tilde{t}\right] + i r \sin\left(\frac{1+\tilde{t}^2}{2}\right) \tilde{t} e^{-ir\tilde{t}}, \tag{67}
\]
\[
F_2(\tilde{t}) = \frac{1}{\sqrt{1+\tilde{t}^2}} \sin\left(\frac{1+\tilde{t}^2}{2}\right) \tilde{t} e^{-ir\tilde{t}}, \tag{68}
\]
\[
F_3(\tilde{t}) = \left[\cos\left(\frac{1+\tilde{t}^2}{2}\right) \tilde{t}\right] - i r \sin\left(\frac{1+\tilde{t}^2}{2}\right) \tilde{t} e^{-ir\tilde{t}}. \tag{69}
\]

### B Solutions using the Laplace transform

In order to solve Eqs. \((26)-(28)\), we rewrite them in the Laplace domain
\[
\frac{d}{dz} \hat{a}_s(z; q) = -g\sqrt{\hat{N}} \hat{a}_s(z; q) \tag{70}
\]
\[-\hat{c}(z, 0; q) + (s + i\Delta)\hat{c}_s(z; q) = g\sqrt{\hat{N}} \hat{a}_s(z; q) + \Omega \hat{b}_s(z; q), \tag{71}\]
\[-\hat{b}(z, 0; q) + s \hat{b}_s(z; q) = -\Omega \hat{c}_s(z; q). \tag{72}\]

From this we can write the closed differential equation for the field amplitude \(\hat{a}_s(z; q)\) in the form
\[
\frac{d\hat{a}_s(z; q)}{dz} = -\Gamma_s \hat{a}_s(z; q) - g\sqrt{\hat{N}} \hat{a}_s(z; q). \tag{73}\]
Here the coefficient $\Gamma_s$ determines a rate of escape of the amplitude along the z-axis and reads

$$\Gamma_s = \frac{g^2 N}{2} \left( \frac{\mu}{s + i\mu \tilde{\Omega}} + \frac{\nu}{s - i\nu \tilde{\Omega}} \right),$$

where the following notations are introduced

$$\tilde{\Omega} = \sqrt{1 + r^2}, \quad \mu = 1 + r, \quad \nu = 1 - r. \quad (75)$$

The inhomogeneous term on the right in Eq. (73) is determined by the initial conditions for the medium state and given by

$$\dot{\alpha}_s(z; q) = \frac{1}{s(s + i\Delta) + \Omega^2} \left[ \Omega \dot{b}(0, z; q) + s \dot{c}(0, z; q) \right].$$

A solution of Eq. (76) reads

$$\hat{a}_s(z; q) = \hat{a}_s(0; q)e^{-\Gamma_s z} - g\sqrt{N} \int_0^z dz' \hat{\alpha}_s(z'; q)e^{-\Gamma_s(z-z')}.$$ \quad (77)

From Eqs. (71)-(72) one can obtain

$$\hat{c}_s(z; q) = g\sqrt{N} \frac{s}{s(s + i\Delta) + \Omega^2} \hat{a}_s(z; q) + \hat{a}_s(z; q), \quad (78)$$

$$\hat{b}_s(z; q) = \frac{1}{s} \left[ \hat{b}(0, z; q) - \Omega \hat{c}_s(z; q) \right].$$ \quad (79)

After the inverse Laplace transformation in Eqs. (77)-(79) one can obtain all the solutions in the form

$$\hat{a}(t; z; q) = \int_0^t dt' \hat{a}_a(t-t'; q) \, D(t', z) \quad (80)$$

$$-g\sqrt{N} \int_0^t dt' \int_0^z dz' \hat{\alpha}(t-t', z-z'; q) \, D(t', z'), \quad \hat{c}(t; z; q) =$$

$$g\sqrt{N} \int_0^t dt' F_3(t-t') \hat{a}(t', z; q) + \hat{a}(t, z; q), \quad (81)$$

$$\hat{b}(t; z; q) = -g\sqrt{N} \int_0^t dt' F_2(t-t') \hat{a}(t', z; q) + \hat{\beta}(z, t; q), \quad (82)$$

where the terms $\hat{a}(t; z; q)$ and $\hat{\beta}(z; t; q)$ are expressed via the initial conditions for the medium coherences in the form

$$\hat{a}(t; z; q) = F_3(t) \hat{c}(0, z; q) + F_2(t) \hat{b}(0, z; q),$$ \quad (83)

$$\hat{\beta}(z; t; q) = -F_2(t) \hat{c}(0, z; q) + F_1(t) \hat{b}(0, z; q),$$ \quad (84)

and the kernel $D(t, z)$ is simply proportional to $G_{aa}$ (see (56)) and is given by

$$D(t, z) = \Omega G_{aa}(t, z).$$ \quad (85)

After some simple transformations one can obtain the solution in the form (51)-(53). At the same time for the numerical computation it is possible to use Eqs. (80)-(82).
Fig. 2. Normalized coherence at $\tilde{z} = 0$ as a function of time for (a) $r = 0$, (b) $r = 0.1$, (c) $r = 0.2$, (d) $r = 1$, (e) $r = 2$, (f) $r = 10$.

Fig. 3. Distribution of the coherence $|\tilde{b}^W(\tilde{f}, \tilde{z})|^2$ in time and space for (a) $r = 0$, (b) $r = 1$ and (c) $r = 2$.

Fig. 4. Normalized distributions of the atomic coherence inside the medium at time $\tilde{t} = \pi$ for (a) $r = 0$, (b) $r = 0.5$, (c) $r = 2$.

Fig. 5. Fraction of signal photons (normalized to the energy of input signal pulse) that have been converted to atomic coherence $b^W$ during writing as a function of the detuning parameter $r$ for $\tilde{L} = 10$ and (a) $\tilde{T}^W = \pi/4$, (b) $\tilde{T}^W = \pi/2$, (c) $\tilde{T}^W = \pi$, (d) $\tilde{T}^W = 2\pi$. 
Fig. 6. Writing process: relative losses associated with leakage (blue curves, dotted lines) and relative total losses (red curves, full lines) (in percent of the input field intensity) at the output of the medium as a function of $\tilde{T}_W$ for $\tilde{L} = 10$ for (a) $r = 0$, (b) $r = 0.5$, (c) $r = 2$.

Fig. 7. Writing process: relative losses due to leakage (blue curves, dotted lines) and relative total losses (red curves, full lines) (in percent of the input field intensity) at the output of the medium as a function of $\tilde{L}$ and (a) $r = 0$, (b) $r = 0.5$ and (c) $r = 2$ for $T_W = \pi/2$ (first row) and $T_W = \pi$ (second row).

Fig. 8. Comparison of relative total losses (in percent of the input field intensity) at the output of the medium as a function of $T_W$ for resonant (blue curves, dotted lines) and detuned (red curves, full lines) cases; $\tilde{L} = 10$, (a) $r = 0.1$, (b) $r = 0.2$, (c) $r = 0.3$, (d) $r = 0.4$. 
Fig. 10. Reading process: field intensity $|a^R(\tilde{T}, \tilde{L})|^2$ at the output of the medium for $\tilde{L} = 10$ and $r = 0$ for backward retrieval with two optimization techniques: (a) total loss minimization ($\tilde{T}^W = 5.5$) and (b) leakage minimization ($\tilde{T}^W = 4.2$).

Fig. 11. Reading process: field intensity $|a^R(\tilde{t}, \tilde{L})|^2$ at the output of the medium for $\tilde{L} = 10$, $\tilde{T}^W = 3$ and $r = 0.5$ for (a) forward and (b) backward propagating retrieval.