

NWCRG
Internet-Draft
Intended status: Informational
Expires: 12 June 2022

S. Yang
CUHK (SZ)
X. Huang
R. W. Yeung
CUHK
J. K. Zao
NCTU
9 December 2021

BATS Coding Scheme for Multi-hop Data Transport
draft-irtf-nwcrg-bats-03

Abstract

BATS code is a class of efficient linear network coding scheme with a matrix generalization of fountain codes as the outer code, and batch-based linear network coding as the inner code. This document describes a baseline BATS coding scheme for communication through multi-hop networks, and discusses the related research issues towards a more sophisticated BATS coding scheme. This document is a product of the Coding for Efficient Network Communications Research Group (NWCRG).

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 12 June 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	3
1.1. Requirements Language	4
2. A Use Case of BATS Coding Scheme	4
2.1. Introduction	5
2.2. Data Delivery Procedures	6
2.2.1. Source Node Data Partitioning and Padding	6
2.2.2. Source Node Outer Code Encoding Procedure	7
2.2.3. Recoding Procedures	9
2.2.4. Destination Node Procedures	10
2.3. Recommendation for the Parameters	10
2.4. Coding Parameters in DDP Packets	11
2.4.1. Coding Parameter Format	11
2.4.2. Coded Packet Format	12
3. BATS Code Specification	13
3.1. Common Parts	13
3.2. Outer Code Encoder	14
3.3. Inner Code Encoder (Recoder)	15
3.4. Outer Decoder	16
4. Research Issues	17
4.1. Coding Design Issues	17
4.2. Protocol Design Issues	18
4.3. Application Related Issues	19
5. IANA Considerations	20
6. Security related Considerations	20
6.1. Preventing Eavesdropping	20
6.2. Countermeasures against Pollution Attacks	21
7. References	22
7.1. Normative References	22
7.2. Informative References	22
Acknowledgments	24
Authors' Addresses	24

1. Introduction

This document specifies a baseline BATS code [Yang14] scheme for data delivery in multi-hop networks, and discusses the related research issues towards a more sophisticated scheme. The BATS code described here includes an outer code and an inner code. The outer code is a matrix generalization of fountain codes (see also the RaptorQ code described in RFC 6330 [RFC6330]), which inherits the advantages of reliability and efficiency and possesses the extra desirable property of being network coding compatible. The inner code, also called recoding, is formed by linear network coding for combating packet loss, improving the multicast efficiency, etc. A detailed design and analysis of BATS codes are provided in the BATS monograph [Yang17].

A BATS coding scheme can be applied in multi-hop networks formed by wireless communication links, which are inherently unreliable due to interference. Existing transport protocols like TCP use end-to-end retransmission, while network protocols such as IP might enable store-and-forward at the relays, so that packet loss would accumulate along the way.

A BATS coding scheme can be used for various data delivery applications like file transmission, video streaming over wireless multi-hop networks, etc. Different from traditional forward error correcting (FEC) schemes that are applied either hop-by-hop or end-to-end, the BATS coding scheme combines the end-to-end coding (the outer code) with certain hop-by-hop coding (the inner code), and hence can potentially achieve better performance.

The baseline coding scheme described here considers a network with multiple communication flows. For each flow, the source node encodes the data for transmission separately. Inside the network, however, it is possible to mix the packets from different flows for recoding. In this document, we describe a simple case where recoding is performed within each flow. Note that the same encoding/decoding scheme described here can be used with different recoding schemes as long as they follow the principle as we illustrate in this document.

The purpose of the baseline BATS coding scheme is twofold. First, it provides researchers and engineers a starting point for developing network communication applications/protocols based on BATS codes. Second, it helps to make the research issues clearer towards a sophisticated BATS code based network protocol. Important research directions include the security issues, congestion control and routing algorithms for BATS codes, etc.

This document is a product of and represents the collaborative work and consensus of the Coding for Efficient Network Communications Research Group (NWCRG); it is not an IETF product and is not an IETF standard.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. A Use Case of BATS Coding Scheme

The BATS coding scheme described in this document can be used, for example, by a Data Delivery Protocol (DDP). Though this document is not about a DDP, we briefly illustrate in this section how a BATS coding scheme is employed by a DDP to make the role of the coding scheme clear. Some terms that will be used in this section are summarized here:

- * DDP: data delivery protocol.
- * DDP packet: the packet formed by a DDP employing a BATS coding scheme.
- * source packet: the packet formed by the data for delivery.
- * outer encoder: the outer code encoder of a BATS code.
- * recoder: the inner code encoder of a BATS code.
- * outer decoder: the outer code decoder of a BATS code.
- * coded packet: the packet generated by the outer code encoder or a recoder.
- * batch: a set of coded packets generated by a BATS coding scheme from the same subset of the source packets.
- * recoded packet: a coded packet generated by a recoder.
- * degree: the number of source packets used to generate a batch by the outer encoder. The degree can be different for different batch.

Other common terms can be found in RFC 8406 [RFC8406].

2.1. Introduction

We describe a data delivery process that involves one source node, one destination node, and multiple intermediate nodes in between. A BATS coding scheme includes an outer code encoder (also called outer encoder), an inner code encoder (also called recoder), and an outer decoder which decodes the outer code and the inner code jointly as illustrated in Figure 1. The functions of the outer encoder, recoder and outer decoder are described below:

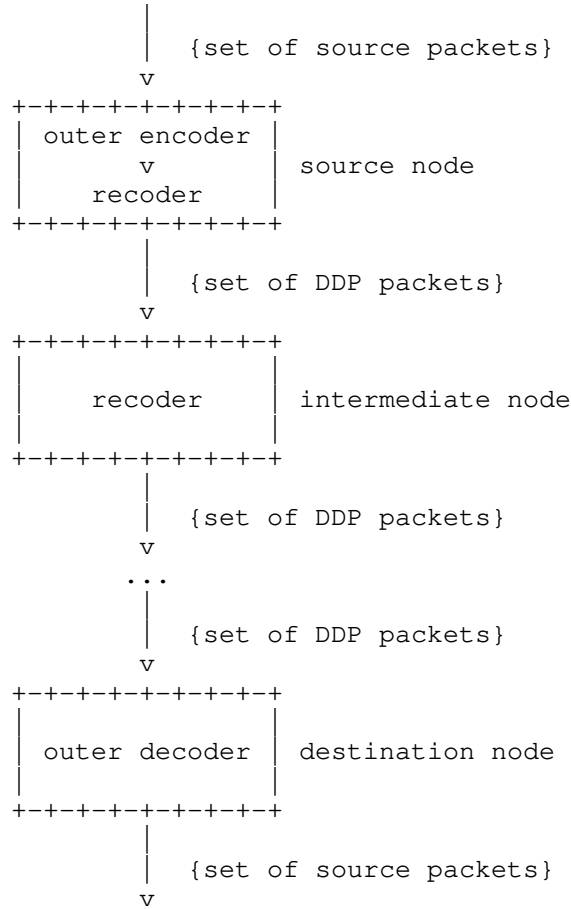


Figure 1: A network model for data delivery.

At the source node, the DDP first processes the data to be delivered into a number of source packets each of the same number of bits (see details in Section 2.2.1), and then provides all the source packets to the outer encoder. The outer encoder will further generate a sequence of batches, each consisting of a fixed number of coded packets (see the description in Section 2.2.2).

Each batch generated at the source node is further processed by the recoder separately. The recoder may generate a number of new coded packets using the existing coded packets of the batch (see the description in Section 2.2.3). After processed by the recoder, the DDP forms and transmits the DDP packets using the coded packets, together with the corresponding batch information.

We assume that a DDP packet is either correctly received or completely erased during the communication. The DDP extracts the coded packets and the corresponding batch information from its received DDP packets. A recoder is employed at an intermediate node that does not need the data, and generates recoded packets for each batch (see the description in Section 2.2.3). The DDP forms and transmits DDP packets using the recoded packets and the corresponding batch information.

The outer decoder is employed at the destination node that needs the data. The DDP extracts the coded packets and the corresponding batch information from its received DDP packets. The outer decoder tries to recover the transmitted data using the received batches (see the description in Section 2.2.4). The DDP sends the decoded data to the application that needs the data.

2.2. Data Delivery Procedures

Suppose that the DDP has F octets of data for transmission. We describe the procedures of one BATS session for transmitting the F octets. There is a limit on F of a single BATS session. If the total data has more than the limit, the data needs to be transmitted using multiple BATS sessions. The limit on F of a single BATS session depends on the coding parameters to be discussed in this section, and will be calculated at the end of Section 2.4.2.

2.2.1. Source Node Data Partitioning and Padding

The DDP first determines the following parameters:

- * Batch size (M): the number of coded packets in a batch generated by the outer encoder.

- * Recoding field size (q): the number of elements in the finite field for recoding. q is 2 or 2^8
- * BATS payload size (T_0): the number of payload octets in a BATS packet, including the coded data and the coefficient vector.

Based on the above parameters, the parameters T , O and K are calculated as follows:

- * O : the number of octets of a coefficient vector, calculated as $O = \text{ceil}(M \cdot \log_2(q)/8)$, which is also called the coefficient vector overhead.
- * T : the number of data octets of a coded packet, calculated as $T = T_0 - O$.
- * K : number of source packets, calculated as $K = \text{floor}(F/T) + 1$.

The data MUST be padded to have $T \cdot K$ octets, which will be partitioned into K source packets $b[0], \dots, b[K-1]$, each of T octets. In our padding scheme, $b[0], \dots, b[K-2]$ are filled with data octets, and $b[K-1]$ is filled with the remaining data octets and padding octets. Let $P = K \cdot T - F$ denote the number of padding octets. We use $b[K-1, 0], \dots, b[K-1, T-P-1]$ to denote the $T-P$ source octets and $b[K-1, T-P], \dots, b[K-1, T-1]$ to denote the P padding octets in $b[K-1]$, respectively. The padding insertion process is shown in Figure 2.

```

Z = T - P
j = 1
v = 1
Let bl be the last source packet b[K-1]
for i = Z, Z+1, ..., T-1 do
    bl[i] = j
    if i+1 >= v+Z do
        j += 1
        v += j

```

Figure 2: Data Padding Insertion Process

2.2.2. Source Node Outer Code Encoding Procedure

The DDP provides the BATS encoder with the following information:

- * Batch size (M): the number of coded packets in a batch.
- * Recoding field size (q): the number of elements in the finite field for recoding.

- * Maximum degree (MAX_DEG): a positive integer that specifies the largest degree can be used.
- * Degree distribution (DD): an unsigned integer array of size MAX_DEG+1. The i -th entry DD[i] is the possibility that i is chosen as the degree, where i is between 0 and MAX_DEG.
- * A sequence of batch IDs (BID) (j , $j = 0, 1, \dots$).
- * Number of source packets (K).
- * Packet size (T): the number of octets in a source packet.
- * Source packets ($b[i]$, $i = 0, 1, \dots, K-1$).

Using this information, the outer encoder generates M coded packets for each batch ID using the following steps to be described in details at Section 3.2:

- * Obtain a degree d by sampling DD. Roughly, the value d is chosen with probability DD[d].
- * Choose d source packets uniformly at random from all the K source packets. It is allowed that a source packet is used by multiple batches.
- * Generate M coded packets using the d source packets.

The DDP receives from the outer encoder a sequence of batches, where the batch with ID j has

- * M coded packets ($x[j,i]$, $i = 0, 1, \dots, M-1$), each containing T_0 octets.

The DDP will use the batches to form DDP packets to be transmitted to other network nodes towards the destination nodes. The DDP MUST deliver with each coded packet with its

- * BID: batch ID.

The DDP MUST deliver the following information to each recoder:

- * M : batch size
- * q : recoding field size.

The DDP MUST deliver the following information to each decoder:

- * M: batch size
- * q: recoding field size
- * K: the number of source packets
- * T: the number of octets in a source packet
- * DD: the degree distribution.

The BID is used by both recoders and decoders. We will illustrate in Section 2.4 that how to embed BID, M, q, and K into DDP packets. The degree distribution DD does not need to be changed frequently. See Section 6 in [Yang17] about how to design a good degree distribution. Once designed, the degree distribution can be shared between the source node and the destination node by the DDP, which is not further discussed here.

2.2.3. Recoding Procedures

Both the source node and the intermediate nodes perform recoding on the batches before transmission. At the source node, the recoder receives the batches from the outer code encoding procedure. At an intermediate node, the DDP receives the DDP packets from the other network nodes. If the DDP choose not to recode, it just forwards the DDP packets it received. Otherwise, the DDP should be able to extract coded packets and the corresponding batch information from these packets.

For a received batch, the DDP determines a positive integer M_r , the number of recoded packets to be transmitted for the batch, and provides the recoder with the following information:

- * the batch size M,
- * the recoding field size q,
- * a number of received coded packets of the same batch, each containing TO octets, and
- * the number of recoded packets to be generated (M_r).

The recoder uses the information provided by the DDP to generate M_r recoded packets, each containing TO octets, to be described in Section 3.3. The DDP uses the M_r recoded packets to form the DDP packets for transmitting.

2.2.4. Destination Node Procedures

A destination node needs the data transmitted by the source node. At the destination node, the DDP receives DDP packets from an intermediate network node, and should be able to extract coded packets and the corresponding batch information from these packets.

The DDP provides the outer decoder (to be described in Section 3.4) with the following information:

- * M: batch size,
- * q: recoding field size,
- * K: the number of source packets
- * T: the number of octets of a source packet
- * A sequence of batches, each of which is formed by a number of coded packets belonging to the same batch, with their corresponding BIDs.

The decoder uses this information to decode the outer code and the inner code jointly and recover the K source packets (see details in Section 3.4). If successful, the decoder returns the recovered K source packets to the DDP, which will use the K source packets to form the F octets data. The recommended padding deletion process is shown as follows:

```
// this procedure returns the number P of padding octets
// at the end of b[K-1]
Let bl be the last decoded source packet b[K-1]
PL = bl[T-1]
if PL == 1 do
    return P = 1
WI = T - 1
while bl[WI] == PL do
    WI = WI - 1
return P = (1 + bl[WI]) * bl[WI] + T - WI - 1
```

Figure 3: Data Padding Deletion Process

2.3. Recommendation for the Parameters

The recommendation for the parameters M and q is shown as follows:

- * When q=2, M=16,32,64,128

- * When $q=256$, $M=4,8,16,32$

It is RECOMMENDED that K is at least 128. The encoder/decoder SHALL support an arbitrary positive integer value less than 2^{16} . However, the BATS coding scheme to be described is not optimized for small K .

2.4. Coding Parameters in DDP Packets

Here we provide an example of embedding the aforementioned BATS coding parameters into the DDP packets which will be used for recoding and decoding. A DDP can form a DDP packet using a coded packet by adding necessary information that can help to deliver the DDP packet to the next node, e.g., the DDP protocol version, addresses and session identifiers. We will not go into the details of formatting these fields in a DDP packet, but focus on how to format the coding parameters and the coded packet in a DDP packet.

2.4.1. Coding Parameter Format

Here we provide an example of using 32 bits (4 octets) to embed the parameters K , M , q , and BID . The 32 bits are separated into three subfields, denoted as K , Mq and BID , respectively, as illustrated in Figure 4.

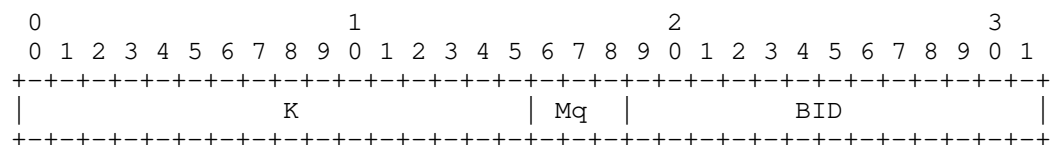


Figure 4: Coding parameter field format.

- * K : 16-bit unsigned integer, specifying the number of source packets of the BATS session.
- * Mq : 3-bit unsigned integer to specify the value of M and q as Table 1.
- * BID : 13-bit unsigned integer, specifying the batch ID of the batch the packet belongs to.

Mq	M	q
000	16	2
010	32	2
100	64	2
110	128	2
001	4	256
011	8	256
101	16	256
111	32	256

Table 1: Values of Mq field

The choice of the coding parameters depends on the computation cost, the network conditions and the expected end-to-end coding performance. Usually, a larger batch size M will have a better coding performance, but higher computation cost for encoding, recoding and decoding. The field size q affects the coefficient vector overhead, and also the computation cost for recoding. Within a BATS session, the BID field should be different for all batches, and hence the maximum number of batches can be generated for the outer encoder is 2^{13} . For different BATS sessions, batches can use the same BID.

2.4.2. Coded Packet Format

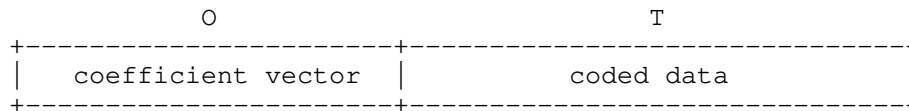


Figure 5: Code packet format in a DDP packet.

A coded packet has $T+O$ octets, where the first O octets contain the coefficient vector and the remaining T octets contain the coded data.

- * coefficient vector: $O = M \cdot \log_2(q)/8$ octets. For the values of M and q in Table 1, O is at most 32 octets when q is 256 and 6 octets when q is 2.
- * coded data: T octets. T should be chosen so that the whole DDP packet is at most PMTU.

Using the above formation, we can calculate the largest file size F for different parameters. For example, when $q=2$ and $M=128$, we have $O = 6$ octets. Counting the 4 octets for embedding the coding parameters, we can choose $T = \text{PMTU} - H - 10$, where H is the header length of a DDP packet. As K can be at most $2^{16}-1$, F can be at most $(\text{PMTU} - H - 10)(2^{16}-1)$ octets. In this case, K/M is about 2^9 and the BID field allows transmitting $2^4 \cdot K/M$ batches.

3. BATS Code Specification

3.1. Common Parts

The T octets of a source packets are treated as a column vector of T elements in $\text{GF}(256)$. The O octets of coefficient vector are treated as a column vector of O elements in $\text{GF}(q)$, where $q=2$ or $q=256$. Linear algebra and matrix operations over finite fields are assumed in this section.

For the two elements of $\text{GF}(2)$, their multiplication corresponds to a logical AND operation and their addition is an logical XOR operation. An element of the field $\text{GF}(256)$ can be represented by a polynomial with binary coefficients and degree lower or equal to 7. The addition between two elements of $\text{GF}(256)$ is defined as the addition of the two binary polynomials. The multiplication between two elements of $\text{GF}(256)$ is the multiplication of the two binary polynomials modulo a certain irreducible polynomial of degree 8, called a primitive polynomial. One example of such a primitive polynomial for $\text{GF}(256)$ is:

$$x^8 + x^4 + x^3 + x^2 + 1$$

A common primitive polynomial should be specified for all the finite field multiplications over $\text{GF}(256)$. Note that a binary polynomial of degree less than 8 can be represented by a binary sequence of 8 bits, i.e., an octet.

Suppose that a pseudorandom number generator `Rand()` which generates an unsigned integer of 32 bits is shared by both encoding and decoding. The pseudorandom generator can be initialized by `Rand_Init(S)` with seed `S`. When `S` is not provided, the pseudorandom generator is initialized arbitrarily. One example of such a pseudorandom generator is defined in RFC 8682 [RFC8682].

A function called `BatchSampler` is used in both encoding and decoding. The function takes two integers `j` and `d` as input, and generates an array `idx` of `d` integers and a `d x M` matrix `G`. The function first initializes the pseudorandom generator with `j`, sample `d` distinct integers from 0 to `K-1` as `idx`, and sample `d*M` integers from 0 to 255 as `G`. See the pseudocode in Figure 6.

```
function BatchSampler(j,d)
    // initialize the pseudorandom generator by seed j.
    Rand_Init(j)
    // sample d distinct integers between 0 and K-1.
    for k = 0, ..., d-1 do
        r = Rand() % K
        while r already exists in idx do
            r = Rand() % K
        idx[k] = r

    // sample d x M matrix
    for r = 0, ..., d-1 do
        for c = 0, ..., M-1 do
            G[r,c] = Rand() % 256

    return idx, G
```

Figure 6: Batch Sampler Function

3.2. Outer Code Encoder

Define a function called `DegreeSampler` that returns an integer `d` using the degree distribution `DD`. We expect that the empirical distribution of the returning `d` converges to `DD(d)` when `d < K`. One design of `DegreeSampler` is illustrated in Figure 7. Note that when `K < MAX_DEG`, the degree value returned by `DegreeSampler` does not exactly follow the distribution `DD`, which however would not affect the practical decoding performance for the outer decoder to be described in Section 3.4.

```

function DegreeSampler(j, DD)
  Let CDF be an array
  CDF[0] = 0
  for i = 1, ..., MAX_DEG do
    CDF[i] = CDF[i-1] + DD[i]
  Rand_Init(j)
  r = Rand() % CDF[MAX_DEG]
  for d = 1, ..., MAX_DEG do
    if r >= CDF[d] do
      return min(d,K)
  return min(MAX_DEG,K)

```

Figure 7: Degree Sampler Function.

Let $b[0], b[1], \dots, b[K-1]$ be the K source packets. A batch with BID j is generated using the following steps.

- * Obtain a degree d by calling DegreeSampler with input j .
- * Obtain idx and $G[j]$ by calling BatchSampler with input j and d .
- * Let $B[j] = (b[\text{idx}[0]], b[\text{idx}[1]], \dots, b[\text{idx}[d-1]])$. Form the batch $X[j] = B[j] * G[j]$, whose dimension is $T \times M$.
- * Form the $TO \times M$ matrix $Xr[j]$, where the first O rows of $Xr[j]$ form the $M \times M$ identity matrix I with entries in $GF(q)$, and the last T rows of $Xr[j]$ is $X[j]$.

See the pseudocode of the batch generating process in Figure 8.

```

function GenBatch(j)
  d = DegreeSampler(j)
  (idx, G) = BatchSampler(j,d)
  B = (b[idx[0]], b[idx[i]], ..., b[idx[d-1]])
  X = B * G
  Xr = [I; X]
  return Xr

```

Figure 8: Batch Generation Function.

3.3. Inner Code Encoder (Recoder)

In general, the inner code of a BATS code comprises (random) linear network coding applied on the coded packets belonging to the same batch. The recoded packets have the same BID. Suppose that coded packets $xr[i]$, $i = 0, 1, \dots, r-1$, which have the same BID j , have been received at an intermediate node, and Mr recoded packets are supposed to be generated. Following traditional random linear

network coding, a recoded packet can be generated by random linear combination: (randomly) choose a sequence of coefficients $c[i]$, $i = 0, 1, \dots, r-1$ from $GF(q)$, and generate $c[0]xr[0]+c[1]xr[1]+\dots+c[r-1]xr[r-1]$ as a recoded packet. This recoding approach, called random linear recoding, achieves good network coding performance for multicast when the batch size is sufficiently large.

For unicast communications in a single path as illustrated in Figure 1, it is not necessary to generate all the Mr recoded packets using random linear combination. Instead, $xr[i]$, $i = 0, 1, \dots, r-1$, are directly used as recoded packets, and $\max(Mr-r, 0)$ recoded packets are generated using linear combinations. Compared with random linear recoding, this recoding approach, called systematic recoding, can reduce both the computation cost and also the recoding latency that accumulates linearly with the number of nodes. Note that the use of systematic recoding may not always achieve the optimal network coding performance as random linear recoding in more complicated communication scenarios that include multiple paths and multiple destination nodes.

3.4. Outer Decoder

The decoder receives a sequence of batches $Yr[j]$, $j = 0, 1, \dots, n-1$, each of which is a TO -row matrix over $GF(256)$. Let $Y[j]$ be the submatrix of the last T rows of $Yr[j]$. When $q = 256$, let $H[j]$ be the first M rows of $Yr[j]$; when $q = 2$, let $H[j]$ be the matrix over $GF(256)$ formed by embedding each bit in the first $M/8$ rows of $Yr[j]$ into $GF(256)$. For successful decoding, we require that the total rank of all the batches is at least K .

The same degree distribution DD used for the outer encoder is supposed to be known by the outer decoder. By calling `DegreeSampler` and `BatchSampler` with input j , we obtain $d[j]$, $idx[j]$ and $G[j]$. According to the encoding and recoding processes described in Section 3.2 and Section 3.3, we have the system of linear equations $Y[j] = B[j]G[j]H[j]$ for each received batch with ID j , where $B[j] = (b[idx[j], 0], b[idx[j], 1], \dots, b[idx[j], d-1])$ is unknown.

We first describe a belief propagation (BP) decoder that can efficiently solve the source packets when a sufficient number of batches have been received. A batch j is said to be decodable if $\text{rank}(G[j]H[j]) = d[j]$ (i.e., the system of linear equations $Y[j] = B[j]G[j]H[j]$ with $B[j]$ as the variable matrix has a unique solution). The BP decoding algorithm has multiple iterations. Each iteration is formed by the following steps:

- * Decoding step: Find a batch j that is decodable. Solve the corresponding system of linear equations $Y[j] = B[j]G[j]H[j]$ and decode $B[j]$.
- * Substitution step: Substitute the decoded source packets into undecodable batches. Suppose that a decoded source packet $b[k]$ is used in generating an undecodable $Y[j]$. The substitution involves 1) removing the entry in $idx[j]$ corresponding to k , 2) removing the row in $G[j]$ corresponding to $b[k]$, and 3) reducing $d[j]$ by 1.

The BP decoder repeats the above steps until no batches are decodable during the decoding step.

When the degree distribution DD in the outer code encoder (see Section 3.2) is properly designed, the BP decoder guarantees a high probability for the recovery of a given fraction of the source packets when K is large. To recover all the source packets, a precode can be applied to the source packets to generate a fraction of redundant packets before applying the outer code encoding. Moreover, when the BP decoder stops which may happen with a high probability when K is relatively small, it is possible to continue with inactivation decoding, where certain source packets are treated inactive so that a similar belief propagation process can be resumed. The reader is referred to RFC 6330 [RFC6330] for the design of a precode with a good inactivation decoding performance.

4. Research Issues

The baseline BATS coding scheme described in Section 2 and Section 3 needs various refinement and complement towards becoming a more sophisticated network communication application. Various related research issues are discussed in this section, but the security related issues are left to Section 6.

4.1. Coding Design Issues

When the number of batches is sufficiently large, the BATS code specification in Section 3 has nearly optimal performance in the sense that the decoding can be successful with a high probability when the total rank of all the batches used for decoding is just slightly larger than the number of source packet K . But when K is small, the degree sampler function in Figure 7 and the BatchSampler function in Figure 6 based on a pseudorandom generator may not sample all the source packets evenly, so that some of the source packets are not well protected. One approach to solve this issue is to generate a deterministic degree sequence when the number of batches is relatively small, and design a special pseudorandom generator that has a good sampling performance when K is small.

There are research issues related to recoding discussed in Section 3.3. One question is how many recoded packets to generate for each batch. Though it is asymptotically optimal when using the same number of recoded packets for all batches, it has been shown that transmitting a different number of recoded packets for different batches can improve the recoding efficiency. The intuition is that for a batch with a lower rank, a smaller number of recoded packets need to be transmitted. This kind of recoding scheme is called adaptive recoding [Yin19].

Packet loss in network communication is usually bursty, which may harm the recoding performance. One way to resolve this issue is to transmit the packets of different batches in a mixed order, which is also called batch interleaving [Yin20]. How to efficiently interleave batches without increasing too much end-to-end latency is a research issue.

Though we only focus on the BATS coding scheme with one source node and one destination node, a BATS coding scheme can be used for multiple source and destination nodes. To benefit from multiple source nodes, we would need different source nodes to generate statistically independent batches. For communicating the same data to multiple destination nodes, which is also called multicast, it is well-known that linear network coding [Li03] achieves the multicast capacity. BATS codes can benefit from network coding due to its inner code, but how to efficiently implement multicast needs further research.

4.2. Protocol Design Issues

The baseline scheme in this document focuses on reliable communication. There are other issues to be considered towards designing a fully functional DDP based on a BATS coding scheme. Here we discuss some network management issues that are closely related to a BATS coding scheme: routing, congestion control and media access control.

The outer code of a BATS code can be regarded as a channel code for the channel induced by the inner code, and hence the network management algorithms should try to maximize the capacity of the channel induced by the inner code. A network utility maximization problem [Dong20] for BATS coding can be applied to study routing, congestion control and media access control jointly. Compared with the network utility maximization for Internet, there are two major differences. First, the network flow rate is not measured by the rate of the raw packets. Instead, a rank based measurement induced by the inner code is applied for BATS coding schemes. Second, due to recoding, the raw packet rate may not be the same for different links

of a flow, i.e., no flow conservation for BATS coding schemes. These differences affect both the objective and the constraints of the utility maximization problem.

Practical congestion control, routing and media access control algorithms for BATS coding schemes deserve more research efforts. Due to the recoding operation, congestion control cannot be only performed end-to-end. The rate of transmitting batches can be controlled end-to-end, but the number of recoded packets generated for a batch must be controlled at the intermediate nodes, which introduces new research issues for congestion control. For routing, the BATS coding scheme is flexible for implementing multi-path data transmission, and different batches can be transmitted on a different path between a source node and a destination node. Under the scenario of BATS coding schemes, media access control can have some different considerations: Retransmission is not necessary, and a reasonably high packet loss rate can be tolerated.

4.3. Application Related Issues

There are more research issues pertaining to different applications. The reliable communication technique provided by BATS codes can be used for a broad range of network communication scenarios. In general, a BATS coding scheme is suitable for data delivery in networks with multiple hops and unreliable links.

One class of typical application scenario is wireless mesh and ad hoc networks [Toh02], including vehicular networks, wireless sensor networks, smart lamppost networks, etc. These networks are characterized by a large number of network devices connected wirelessly with each other without a centralized network infrastructure. A BATS coding scheme is suitable for high data load delivery in such networks without the requirement that the point-to-point/one-hop communication is highly reliable. Therefore, employing a BATS coding scheme can provide more freedom for media access control, including power control, and physical-layer design so that the overall network throughput can be improved.

Another typical application scenario of BATS coding schemes is underwater acoustic networks [Sprea19], where the propagation delay of acoustic waves in underwater can be as long as several seconds. Due to the long delay, feedback based mechanisms become inefficient. Moreover, point-to-point/one-hop underwater acoustic communication (for both the forward and reverse directions) is highly unreliable. Due to these reasons, traditional networking techniques developed for radio and wireline networks cannot be directly applied to underwater networks. As a BATS coding scheme does not rely on the feedback for reliability communication and can tolerate highly unreliable links, it makes a good candidate for developing data delivery protocols for underwater acoustic networks.

Last but not least, due to its capability of performing multi-source multi-destination communications, a BATS coding scheme can be applied in various content distribution scenarios. For example, a BATS coding scheme can be a candidate for the erasure code used in the liquid data networking framework [Byers20] of CCN (content centric networking), and provides the extra benefit of network coding [Zhang16].

5. IANA Considerations

This memo includes no request to IANA.

6. Security related Considerations

Subsuming both random linear network codes (RLNC) and fountain codes, BATS codes naturally inherit both their desirable security capability of preventing eavesdropping, as well as their vulnerability towards pollution attacks. In this section, we discuss some related research issues.

6.1. Preventing Eavesdropping

Suppose that an eavesdropper obtains a batch where the degree value d is strictly larger than the batch size M . Even the eavesdropper has all the related encoding information, the system of linear equations related to this batch does not have a unique solution, and the probability that the eavesdropper can guess the d source packets used for encoding the batch correctly is $2^{-(d-M)T} \geq 2^{-T}$ (see also [Bhattad05]). When inactivation decoding is applied, we can design the degree distribution DD so that the smallest degree is $M+1$, and hence prevent the eavesdropper from decoding source packets from individual batches.

If we allow the eavesdropper to collect multiple batches and use inactivation decoding, the same security holds if the total rank of all the batches collected by the eavesdropper is less than the number of source packet. Therefore, if the DDP can manage to restrict the eavesdropper from collecting a sufficiently number of coded packets, the native security of BATS code is effective when T is sufficiently large. Here by native security, we mean the security protection provided by the BATS coding scheme without extra enhancement.

If the eavesdropper can collect a sufficient number of coded packets for correctly decoding, the native security of BATS code is ineffective. One solution in this case is to encrypt the whole data before using the BATS code scheme. Better schemes are desired towards reducing the computation cost of the whole data encryption solution. This is a research issue that depends on specific BATS code schemes, and will not be further discussed here.

The threat exists for eavesdropping on the initial encoding process, which takes place at the encoding nodes. In these nodes, the transported data are presented in plain text and can be read along their transfer paths. Hence, information isolation between the encoding process and all other user processes running on the source node MUST be assured.

In addition, the authenticity and trustworthiness of the encoding, recoding and decoding program running on all the nodes MUST be attested by a trusted authority. Such a measure is also necessary in countering pollution attacks.

6.2. Countermeasures against Pollution Attacks

Like all network codes, BATS codes are vulnerable to pollution attacks. In these attacks, one or more compromised coding node(s) can pollute the coded messages by injecting forged packets into the network and thus prevent the receivers from recovering the transported data correctly.

The research community has long been investigating the use of various signature schemes (including homomorphic signatures) to identify the forged packets and stall the attacks (see [Zhao07], [Yu08], [Agrawal09]). However, these countermeasures are regarded as being too computationally expensive to be employed in broadband communications. Hence, a system-level approach based on Trusted Computing [TC-Wikipedia] is proposed as a practical alternative to protect BATS codes against pollution attacks. This Trusted Computing based protection consists of the following countermeasures:

1. Attestation and Validation of all BATS encoding, recoding and decoding nodes in the network. Remote attestation and repetitive validation of the identity and capability of these node based on valid public key certificates with proper authorization MUST be a pre-requisite for admitting these nodes to a network and permitting them to remain on that network.
2. Attestation of all encoding, recoding and decoding programs used in the coding nodes. All programs used to perform the BATS encoding, recoding and decoding processes MUST be remotely attested before they are permitted to run on any of the coding nodes. Reloading or alteration of programs MUST NOT be permitted during an encoding session. Programs MUST be attested or validated again when they are executed in new execution environments instantiated even in the same node.
3. Origin Authentication of all coded messages using network level security protocols such as IPsec or Peer Authentication over session-based communication using transport level security protocols such as TLS/DTLS MUST be employed in order to provide Message Integrity or Origin Authentication to every coded packet sent through the coding network.

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC8406] Adamson, B., Adjih, C., Bilbao, J., Firoiu, V., Fitzek, F., Ghanem, S., Lochin, E., Masucci, A., Montpetit, M-J., Pedersen, M., Peralta, G., Roca, V., Ed., Saxena, P., and S. Sivakumar, "Taxonomy of Coding Techniques for Efficient Network Communications", RFC 8406, DOI 10.17487/RFC8406, June 2018, <<https://www.rfc-editor.org/info/rfc8406>>.
- [RFC8682] Saito, M., Matsumoto, M., Roca, V., Ed., and E. Baccelli, "TinyMT32 Pseudorandom Number Generator (PRNG)", RFC 8682, DOI 10.17487/RFC8682, January 2020, <<https://www.rfc-editor.org/info/rfc8682>>.

7.2. Informative References

- [Agrawal09] Agrawal, S. and D. Boneh, "Homomorphic MACs: MAC-based integrity for network coding", International Conference on Applied Cryptography and Network Security , 2009.
- [Bhattad05] Bhattad, K. and K.R. Narayanan, "Weakly Secure Network Coding", ISIT , 2007.
- [Byers20] Byers, J.W. and M. Luby, "Liquid Data Networking", ICN , 2020.
- [Dong20] Dong, Y., Jin, S., Yang, S., and H.H.F. Yin, "Network Utility Maximization for BATS Code enabled Multihop Wireless Networks", ICC , 2020.
- [Li03] Li, S.-Y.R., Yeung, R.W., and N. Cai, "Linear Network Coding", IEEE Transactions on Information Theory , 2003.
- [RFC6330] Luby, M., Shokrollahi, A., Watson, M., Stockhammer, T., and L. Minder, "RaptorQ Forward Error Correction Scheme for Object Delivery", RFC 6330, DOI 10.17487/RFC6330, August 2011, <<https://www.rfc-editor.org/info/rfc6330>>.
- [Sprea19] Sprea, N., Bashir, M., Truhachev, D., Srinivas, K.V., Schlegel, C., and C. Claudio Sacchi, "BATS Coding for Underwater Acoustic Communication Networks", OCEANS , 2019.
- [TC-Wikipedia] "Trusted Computing",
Wikipedia https://en.wikipedia.org/wiki/Trusted_Computing.
- [Toh02] Toh, C.K., "Ad Hoc Mobile Wireless Networks", Prentice Hall Publishers , 2002.
- [Yang14] Yang, S. and R.W. Yeung, "Batched Sparse Codes", IEEE Transactions on Information Theory 60(9), 5322–5346, 2014.
- [Yang17] Yang, S. and R.W. Yeung, "BATS Codes: Theory and Practice", Morgan & Claypool Publishers , 2017.
- [Yin19] Yin, H.H.F., Tang, B., Ng, K.H., Yang, S., Wang, X., and Q. Zhou, "A Unified Adaptive Recoding Framework for Batched Network Coding", ISIT , 2019.
- [Yin20] Yin, H.H.F., Yeung, R.W., and S. Yang, "A Protocol Design Paradigm for Batched Sparse Codes", Entropy , 2020.

- [Yu08] Yu, Z., Wei, Y., Ramkumar, B., and Y. Guan, "An Efficient Signature-Based Scheme for Securing Network Coding Against Pollution Attacks", INFOCOM , 2008.
- [Zhang16] Zhang, G. and Z. Xu, "Combing CCN with network coding: An architectural perspective", Computer Networks , 2016.
- [Zhao07] Zhao, F., Kalker, T., Medard, M., and K.J. Han, "Signatures for content distribution with network coding", ISIT , 2007.

Acknowledgments

The authors would like to thank the NWCRG chairs, Vincent Roca (our shepherd) and Marie-Jose Montpetit; and all those who provided comments -- namely (in alphabetical order), Emmanuel Lochin, David Oran, and Colin Perkins.

Authors' Addresses

Shenghao Yang
CUHK(SZ)
Shenzhen
Guangdong,
China

Phone: +86 755 8427 3827
Email: shyang@cuhk.edu.cn

Xuan Huang
CUHK
Hong Kong
Hong Kong SAR,
China

Phone: +852 3943 8375
Email: 1155136647@link.cuhk.edu.hk

Raymond W. Yeung
CUHK
Hong Kong
Hong Kong SAR,
China

Phone: +852 3943 8375
Email: whyeung@ie.cuhk.edu.hk

John K. Zao
NCTU
Hsinchu
Taiwan,
China

Email: jkzao@ieee.org

NWCRG
Internet-Draft
Intended status: Informational
Expires: 26 August 2022

N. Kuhn
CNES
E. Lochin
ENAC
F. Michel
UCLouvain
M. Welzl
University of Oslo
22 February 2022

Coding and congestion control in transport
draft-irtf-nwcrg-coding-and-congestion-12

Abstract

Forward Erasure Correction (FEC) is a reliability mechanism that is distinct and separate from the retransmission logic in reliable transfer protocols such as TCP. FEC coding can help deal with losses at the end of transfers or with networks having non-congestion losses. However, FEC coding mechanisms should not hide congestion signals. This memo offers a discussion of how FEC coding and congestion control can coexist. Another objective is to encourage the research community to also consider congestion control aspects when proposing and comparing FEC coding solutions in communication systems.

This document is the product of the Coding for Efficient Network Communications Research Group (NWCRG). The scope of the document is end-to-end communications: FEC coding for tunnels is out-of-the scope of the document.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 26 August 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	3
2. Context	4
2.1. Fairness, Quantifying and Limiting Harm, and Policy Concerns	4
2.2. Separate channels, separate entities	5
2.3. Relation between transport layer and application requirements	7
2.4. Scope of the document concerning transport multipath and multi-streams applications	8
2.5. Types of coding	9
3. FEC above the transport	10
3.1. Fairness and impact on non-coded flows	11
3.2. Congestion control and recovered symbols	11
3.3. Interactions between congestion control and coding rates	11
3.4. On useless repair symbols	12
3.5. On partial ordering at FEC level	12
3.6. On partial reliability at FEC level	12
3.7. On multipath transport and FEC mechanism	12
4. FEC within the transport	12
4.1. Fairness and impact on non-coded flows	14
4.2. Interactions between congestion control and coding rates	14
4.3. On useless repair symbols	14
4.4. On partial ordering at FEC and/or transport level	15
4.5. On partial reliability at FEC level	15
4.6. On transport multipath and subpath FEC coding rate	15
5. FEC below the transport	15
5.1. Fairness and impact on non-coded flows	17
5.2. Congestion control and recovered symbols	17
5.3. Interactions between congestion control and coding rates	17

5.4. On useless repair symbols	17
5.5. On partial ordering at FEC level with in-order delivery transport	17
5.6. On partial reliability at FEC level	18
5.7. FEC not aware of transport multipath	18
6. Research recommendations and questions	18
6.1. Activities related to congestion control and coding . . .	18
6.2. Open research questions	18
6.2.1. Parameter derivation	19
6.2.2. New signaling methods and fairness	19
6.3. Recommendations and advice for evaluating coding mechanisms	20
7. Acknowledgements	20
8. IANA Considerations	20
9. Security Considerations	20
10. Informative References	21
Authors' Addresses	23

1. Introduction

There are cases where deploying FEC coding improves the performance of a transmission. As an example, it may take time for a sender to detect transfer tail losses (losses that occur at the end of a transfer, where, e.g., TCP obtains no more ACKs that would enable it to quickly repair the loss via retransmission). Allowing the receiver to recover such losses instead of having to rely on a retransmission could improve the experience of applications using short flows. Another example is a network where non-congestion losses are persistent and prevent a sender from exploiting the link capacity.

Coding and the loss detection of congestion controls are two distinct and separate reliability mechanisms that is distinct and separate from the loss detection of congestion controls. Since FEC coding repairs losses, blindly applying FEC may easily lead to an implementation that also hides a congestion signal from the sender. It is important to ensure that such information hiding does not occur, because loss may be the only congestion signal available to the sender (e.g. TCP [RFC5681]).

FEC coding and congestion control can be seen as two separate channels. In practice, implementations may mix the signals that are exchanged on these channels. This memo offers a discussion of how FEC coding and congestion control coexist. Another objective is to encourage the research community also to consider congestion control aspects when proposing and comparing FEC coding solutions in communication systems. This document does not aim at proposing guidelines for characterizing FEC coding solutions.

We consider three architectures for end-to-end unicast data transfer:

- * with FEC coding in the application (above the transport) (Section 3),
- * within the transport (Section 4), or
- * directly below the transport (Section 5).

A typical scenario for the considerations in this document is a client browsing the web or watching a live video.

This document represents the collaborative work and consensus of the Coding for Efficient Network Communications Research Group (NWCRCG); it is not an IETF product and is not a standard. The document follows the terminology proposed in the taxonomy document [RFC8406].

2. Context

2.1. Fairness, Quantifying and Limiting Harm, and Policy Concerns

Traffic from or to different end users may share various types of bottlenecks. When such a shared bottleneck does not implement some form of flow protection, the share of the available capacity between single flows can help assess when one flow starves the other.

As one example, for residential accesses, the data rate can be guaranteed for the customer premises equipment, but not necessarily for the end user. The quality of service that guarantees fairness between the different clients can be seen as a policy concern [I-D.briscoe-tsvarea-fair].

While past efforts have focused on achieving fairness, quantifying and limiting harm caused by new algorithms (or algorithms with coding) is more practical [BEYONDJAIN]. This document considers fairness as the impact of the addition of coded flows on non-coded flows when they share the same bottleneck. It is assumed that the non-coded flows respond to congestion signals from the network. This document does not contribute to the definition of fairness at a wider scale.

2.2. Separate channels, separate entities

Figure 1 and Figure 2 present the notations that will be used in this document and introduces the Forward Erasure Correction (FEC) and Congestion Control (CC) channels. The Forward Erasure Correction channel carries repair symbols (from the sender to the receiver) and information from the receiver to the sender (e.g. signaling which symbols have been recovered, loss rate prior and/or after decoding, etc.). The Congestion Control channel carries network packets from a sender to a receiver, and packets signaling information about the network (number of packets received vs. lost, Explicit Congestion Notification (ECN) [RFC3168] marks, etc.) from the receiver to the sender. The network packets that are sent by the Congestion Control channel may be composed of source packets and/or repair symbols.

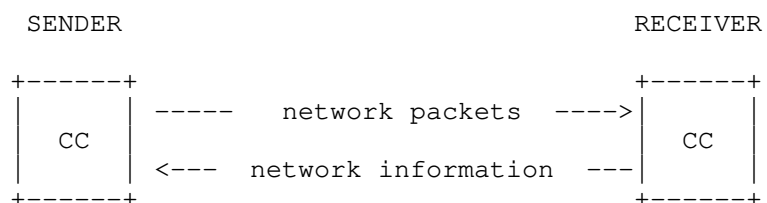


Figure 1: Congestion Control (CC) channel

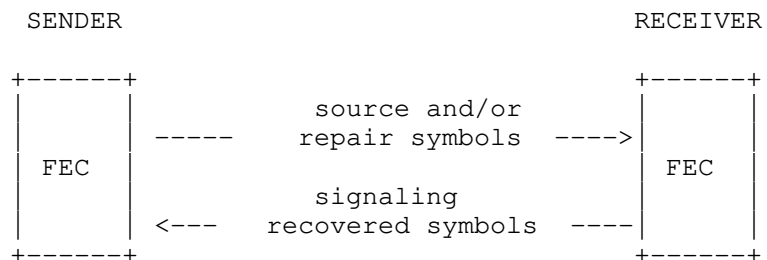


Figure 2: Forward Erasure Correction (FEC) channel

Inside a host, the CC and FEC entities can be regarded as conceptually separate:

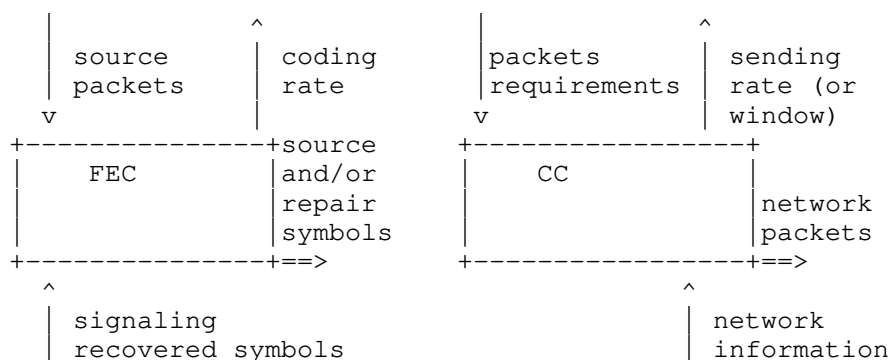


Figure 3: Separate entities (sender-side)

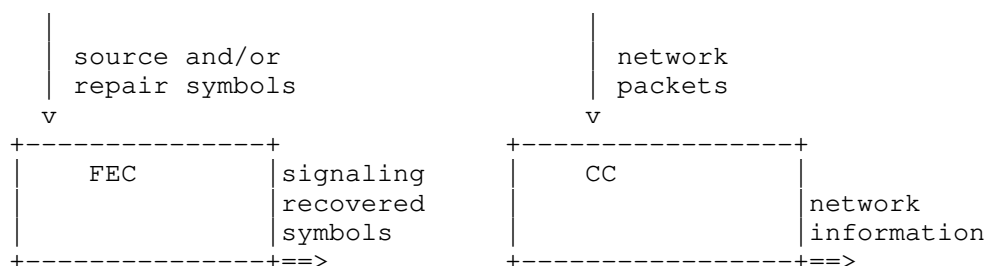


Figure 4: Separate entities (receiver-side)

Figure 3 and Figure 4 provide more details than Figure 1 and Figure 2. Some elements are introduced:

- * 'network information' (input control plane for the transport including CC): refers not only to the network information that is explicitly signaled from the receiver, but all the information a congestion control obtains from a network.
- * 'requirements' (input control plane for the transport including CC): refers to application requirements such as upper/lower rate bounds, periods of quiescence, or a priority.
- * 'sending rate (or window)' (output control plane for the transport including CC): refers to the rate at which a congestion control decides to transmit packets based on 'network information'.
- * 'signaling recovered symbols' (input control plane for the FEC): refers to the information a FEC sender can obtain from a FEC receiver about the performance of the FEC solution as seen by the receiver.

- * 'coding rate' (output control plane for the FEC): refers to the coding rate that is used by the FEC solution (i.e. proportion of transmitted symbols that carry useful data).
- * 'network packets' (output data plane for the CC): refers to the data that is transmitted by a CC sender to a CC receiver. The network packets may contain source and/or repair symbols.
- * 'source and/or repair symbols' (data plane for the FEC): refers to the data that is transmitted by a FEC sender to a FEC receiver. The sender can decide to send source symbols only (meaning that the coding rate is 0), repair symbols only (if the solution decides not to send the original source symbols) or a mix of both.

The inputs to FEC (incoming data packets without repair symbols, and signaling from the receiver about losses and/or recovered symbols) are distinct from the inputs to CC. The latter calculates a sending rate or window from network information, and it takes the packet to send as input, sometimes along with application requirements such as upper/lower rate bounds, periods of quiescence, or a priority. It is not clear that the ACK signals feeding into a congestion control algorithm are useful to FEC in their raw form, and vice versa - information about recovered blocks may be quite irrelevant to a CC algorithm.

2.3. Relation between transport layer and application requirements

The choice of the adequate transport layer may be related to application requirements and the services offered by a transport protocol [RFC8095]:

- * The transport layer may implement a retransmission mechanism to guarantee the reliability of a data transfer (e.g. TCP). Depending on how the FEC and CC functions are scheduled (FEC above CC (Section 3), FEC in CC (Section 4), FEC below CC (Section 5)), the impact of reliable transport on the FEC reliability mechanisms is different.

The transport layer may provide an unreliable transport service (e.g. UDP or DCCP [RFC4340]) or a partially reliable transport service (e.g. SCTP with the partial reliability extension [RFC3758] or QUIC with the unreliable datagram extension [I-D.ietf-quic-datagram]). Depending on the amount of redundancy and network conditions, there could be cases where it becomes impossible to carry traffic. This is further discussed in Section 3 where "FEC above CC" case is assessed and in Section 4 and in Section 5 where "FEC in CC" and "FEC below CC" are assessed.

2.4. Scope of the document concerning transport multipath and multi-streams applications

The application layer can be composed of several streams above FEC and transport layers instances. The transport layer can exploit a multipath mechanism. The different streams could exploit different paths between the sender and the receiver. Moreover, a single-stream application could also exploit a multipath transport mechanism. This section describes what is in the scope of this document in regards with multi-streams applications and multipath transport protocols.

The different combinations between multi-stream applications and multipath transport are the following: (1) one application layer stream as input packets above a combination of FEC and multipath (Mpath) transport layers (Figure 5), and (2) multiple application layer streams as input packets above a combination of FEC and multipath (Mpath) or single path (Spath) transport layers (Figure 6). This document further details cases I (in Section 3.7), II (in Section 4.6) and III (in Section 5.7) illustrated in Figure 5. Cases IV, V and VI of Figure 6 are related to how multiple streams are managed by a single transport or FEC layer: this does not directly concerns the interaction between FEC and the transport and is out of the scope of this document.

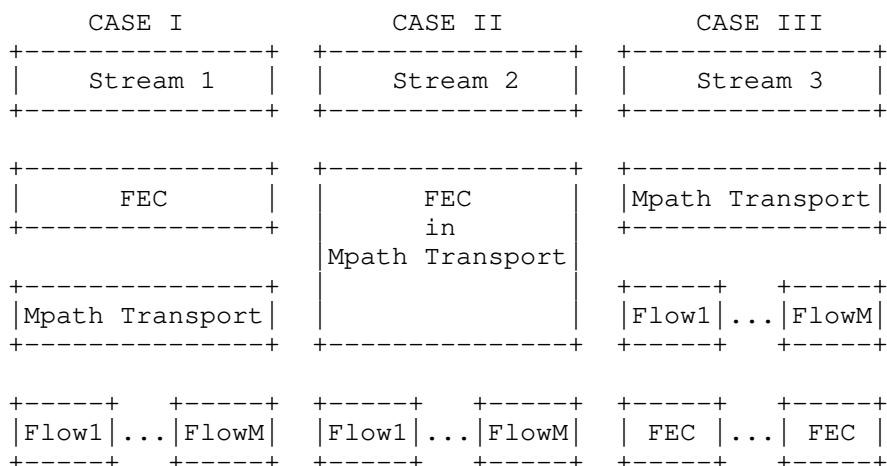


Figure 5: Transport multipath and single stream applications - in the scope of the document

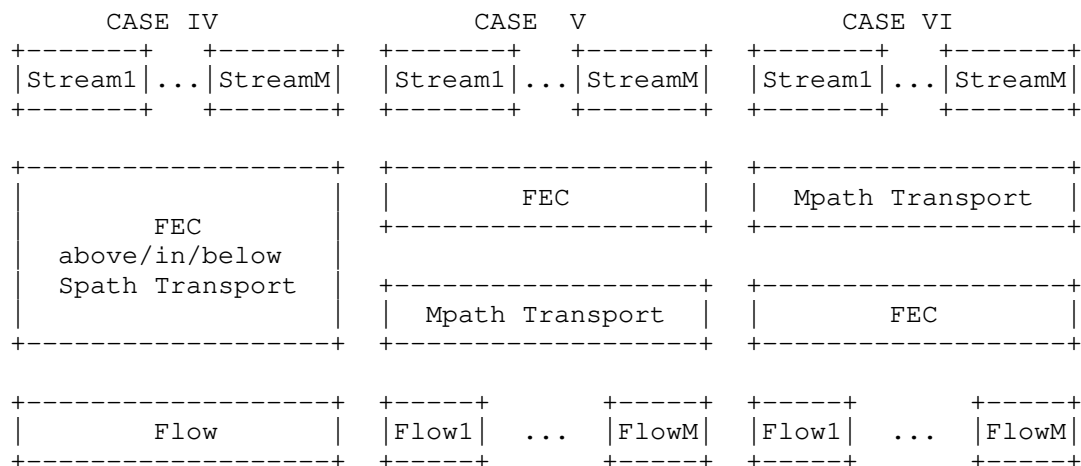


Figure 6: Transport single path, transport multipath and multi-stream applications - out of the scope of the document

2.5. Types of coding

[RFC8406] summarizes recommended terminology for Network Coding concepts and constructs. In particular, the document identifies the following coding types (among many others):

- * **Block Coding:** Coding technique where the input Flow must first be segmented into a sequence of blocks; FEC encoding and decoding are performed independently on a per-block basis.
- * **Sliding Window Coding:** general class of coding techniques that rely on a sliding encoding window.

The decoding scheme may not be able to decode all the symbols. The chance of decoding the erased packets depends on the size of the encoding window, the coding rate and the distribution of erasure in the transmission channel. The FEC channel may let the client transmit information related to the need of supplementary symbols to adapt the level of reliability. Partial and full reliability could be envisioned.

- * **Full reliability:** The receiver may hold symbols until the decoding of source symbols is possible. In particular, if the codec does not enable a subset of the system to be inverted, the receiver would have to wait for a certain minimum amount of repair packets before it can recover all the source symbols.

- * **Partial reliability:** The receiver cannot deliver source symbols that could not have been decoded to the upper layer. For a fixed size of encoding window (for Sliding Window Coding) or of blocks (for Block Coding) containing the source symbols, increasing the amount of repair symbols would increase the chances of recovering the erased symbols. However, this would impact on memory requirements, on the cost of encoding and decoding processes and on the network overhead.

3. FEC above the transport

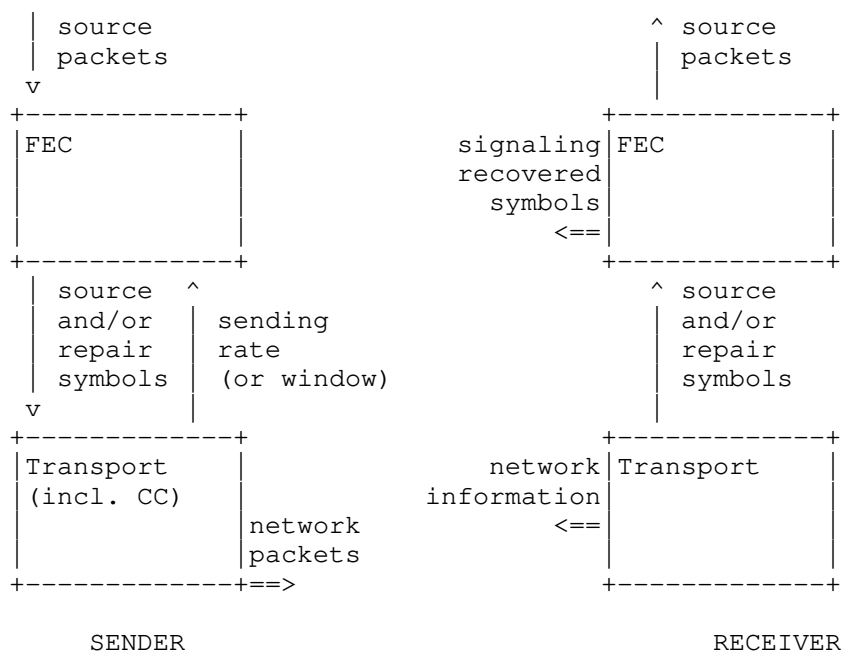


Figure 7: FEC above the transport

Figure 7 presents an architecture where FEC operates on top of the transport.

The advantage of this approach is that the FEC overhead does not contribute to congestion in the network when congestion control is implemented at the transport layer, because the repair symbols are sent following the congestion window or rate determined by the CC mechanism. This can result in improved quality of experience for latency sensitive applications such as Voice over IP (VoIP) or any not-fully reliable services.

This approach requires that the transport protocol does not implement a fully reliable in-order data transfer service (e.g., like TCP). QUIC with unreliable datagram extension [I-D.ietf-quic-datagram] is an example of a protocol for which this is relevant. In cases where the partially reliable transport is blocked and a fall-back to a reliable transport is proposed, there is a risk for bad interactions between reliability at the transport level and coding schemes. For reliable transfers, coding usage does not guarantee better performance; instead, it would mainly reduce goodput.

3.1. Fairness and impact on non-coded flows

The addition of coding within the flow does not influence the interaction between coded and non-coded flows. This interaction would mainly depend on the congestion controls associated with each flow.

3.2. Congestion control and recovered symbols

The congestion control mechanism receives network packets and may not be able to differentiate repair symbols from actual source ones. This differentiation requires a transport protocol providing more than the services described in [RFC8095], in particular specifically indicating what information has been repaired. The relevance of adding coding at the application layer is related to the needs of the application. For real-time applications using an unreliable or partially reliable transport, this approach may reduce the number of losses perceived by the application.

3.3. Interactions between congestion control and coding rates

The coding rate applied at the application layer mainly depends on the available rate or congestion window given by the congestion control underneath. The coding rate could be adapted to avoid adding overhead when the minimum required data rate of the application is not provided by the congestion control underneath. When the congestion control allows sending faster than the application needs, adding coding can reduce packet losses and improve the quality of experience (provided that an unreliable or partially reliable transport is used).

3.4. On useless repair symbols

The only case where adding useless repair symbols does not obviously result in reduced goodput is when the application rate is limited (e.g., VoIP traffic). In this case, useless repair symbols would only impact the amount of data generated in the network. Extra data in the network can, however, increase the likelihood of increasing delay and/or packet loss, which could provoke a congestion control reaction that would degrade goodput.

3.5. On partial ordering at FEC level

Irrespective of the transport protocol, a FEC mechanism does not require to implement a reordering mechanism if the application does not need it. However, if the application needs in-order delivery of packets, a reordering mechanism at the receiver is required.

3.6. On partial reliability at FEC level

The application may require partial reliability. In this case, the coding rate of a FEC mechanism could be adapted based on inputs from the application and the trade-off between latency and packet loss. Partial reliability impacts the type of FEC and type of codec that can be used, such as discussed in Section 2.5.

3.7. On multipath transport and FEC mechanism

Whether the transport protocol exploits multiple paths or not does not have an impact on the FEC mechanism.

4. FEC within the transport

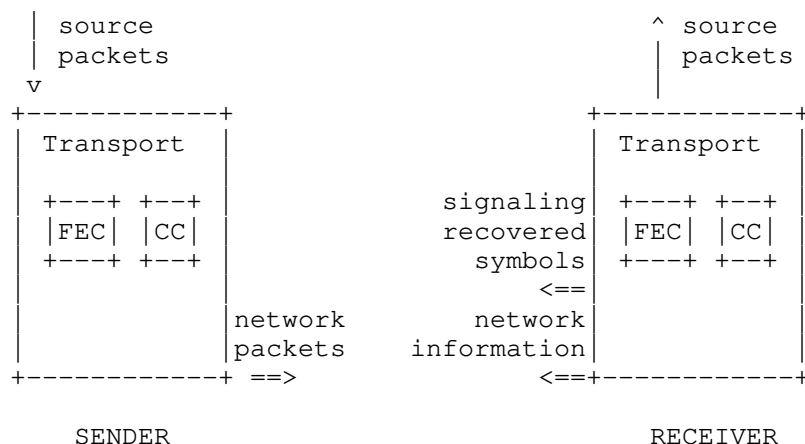


Figure 8: FEC in the transport

Figure 8 presents an architecture where FEC operates within the transport. The repair symbols are sent within what the congestion window or calculated rate allows, such as in [CTCP].

The advantage of this approach is that it allows a joint optimization of CC and FEC. Moreover, the transmission of repair symbols does not add congestion in potentially congested networks but helps repair lost packets (such as tail losses). This joint optimization is the key to prevent flows to consume the whole available capacity. The amount of repair traffic injected should not lead to congestion. As denoted in [I-D.singh-rmcat-adaptive-fec], an increase of the repair ratio should be done conjointly with a decrease of the source sending rate.

The drawback of this approach is that it may require specific signaling and transport services that may not be described in [RFC8095]. Therefore, development and maintenance may require specific efforts at both transport and coding level and the design of the solution may end up being complex to suit different deployment needs.

For reliable transfers, including redundancy reduces goodput for long transfers but the amount of repair symbols can be adapted, e.g. depending on the congestion window size. There is a trade-off between 1) the capacity that could have been exploited by application data instead of transmitting source packets, and 2) the benefits derived from transmitting repair symbols (e.g. unlocking the receive buffer if it is limiting). The coding ratio needs to be carefully designed. For small files, sending repair symbols when there is no more data to transmit could help to reduce the transfer time. Sending repair symbols can avoid the silence period between the transmission of the last packet in the send buffer and 1) firing a retransmission of lost packets, or 2) the transmission of new packets.

Examples of the solution could be to add a given percentage of the congestion window or rate as supplementary symbols, or to send a fixed amount of repair symbols at a fixed rate. The redundancy flow can be decorrelated from the congestion control that manages source packets: a separate congestion control entity could be introduced to manage the amount of recovered symbols to transmit on the FEC channel. The separate congestion control instances could be made to work together while adhering to priorities, as in coupled congestion control for RTP media [RFC8699] in case all traffic can be assumed to take the same path, or otherwise with a multipath congestion window coupling mechanism as in Multipath TCP [RFC6356]. Another possibility would be to exploit a lower than best-effort congestion control [RFC6297] for repair symbols.

4.1. Fairness and impact on non-coded flows

Specific interaction between congestion controls and coding schemes can be proposed (see Section 4.2 and Section 4.3). If no specific interaction is introduced, the coding scheme may hide congestion losses from the congestion controller and the description of Section 5 may apply.

4.2. Interactions between congestion control and coding rates

The receiver can differentiate between source packets and repair symbols. The receiver may indicate both the number of source packets received and repair symbols that were actually useful in the recovery process of packets. The congestion control at the sender can then exploit this information to tune congestion control behavior.

There is an important flexibility in the trade-off, inherent to the use of coding, between (1) reducing goodput when useless repair symbols are transmitted and (2) helping to recover from losses earlier than with retransmissions. The receiver may indicate to the sender the number of packets that have been received or recovered. The sender may use this information to tune the coding ratio. For example, coupling an increased transmission rate with an increasing or decreasing coding rate could be envisioned. A server may use a decreasing coding rate as a probe of the channel capacity and adapt the congestion control transmission rate.

4.3. On useless repair symbols

The sender may exploit the information given by the receiver to reduce the number of useless repair symbols, and improve goodput.

4.4. On partial ordering at FEC and/or transport level

The application may require in-order delivery of packets. In this case, both FEC and transport layer mechanisms should guarantee that packets are delivered in order. If partial ordering is requested by the application, both the FEC and transport could relax the constraints related to in-order delivery: partial ordering impacts both the congestion control and the type of FEC and type of codec that can be used, mostly at the receiver that may need to implement partial reordering.

4.5. On partial reliability at FEC level

The application may require partial reliability. The reliability offered by FEC may be sufficient, with no retransmission required. This depends on application needs and the trade-off between latency and loss. Partial reliability impacts the type of FEC and type of codec that can be used, such as discussed in Section 2.5.

4.6. On transport multipath and subpath FEC coding rate

The sender may adapt the coding rate of each of the single subpaths, whether the congestion control is coupled or not. There is an important flexibility on how the coding rate is tuned depending on the characteristics of each subpath.

5. FEC below the transport

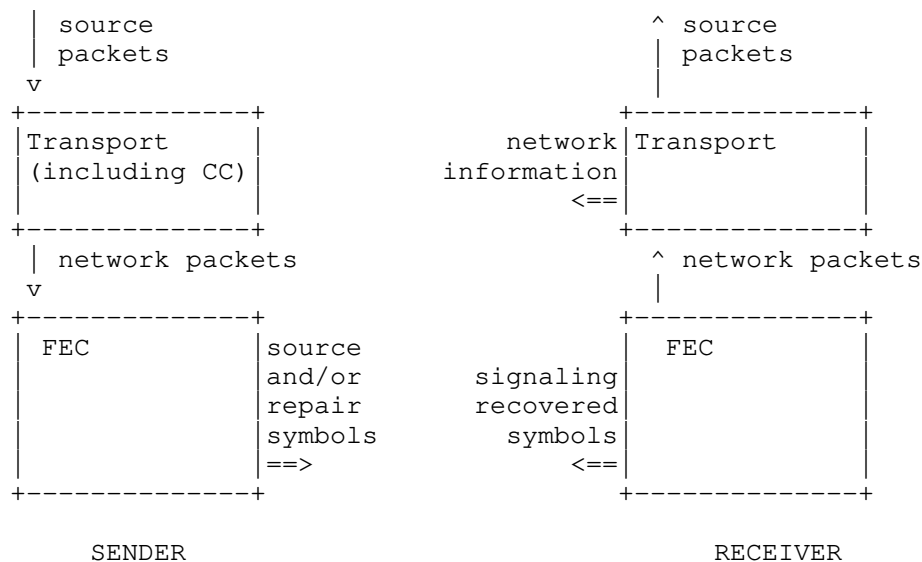


Figure 9: FEC below the transport

Figure 9 presents an architecture where FEC is applied end-to-end below the transport layer, but above the link layer. Note that it is common to apply FEC at the link layer on one or more of the links that make up the end-to-end path. The application of FEC at the link layer contributes to the total capacity that a link exposes to upper layers, but may not be visible to either the end-to-end sender or receiver, if the end-to-end sender and receiver are separated by more than one link, and therefore is out of scope for this document. This includes the use of FEC on top of a link layer in scenarios where the link is known by configuration. In the scenario considered here, the repair symbols are not visible to the end-to-end congestion controller and may be sent on top of what is allowed by the congestion control.

Including redundancy adds traffic without reducing goodput but incurs potential fairness issues. The effective bit-rate is higher than the CC's computed fair share due to the transmission of repair symbols, and losses are hidden from the transport. This may cause a problem for loss-based congestion detection, but it is not a problem for delay-based congestion detection.

The advantage of this approach is that it can result in performance gains when there are persistent transmission losses along the path.

The drawback of this approach is that it can induce congestion in already congested networks. The coding ratio needs to be carefully designed.

Examples of the solution could be to add a given percentage of the congestion window or rate as supplementary symbols, or to send a fixed amount of repair symbols at a fixed rate. The redundancy flow can be decorrelated from the congestion control that manages source packets: a separate congestion control entity could be introduced to manage the amount of recovered symbols to transmit on the FEC channel. The separate congestion control instances could be made to work together while adhering to priorities, as in coupled congestion control for RTP media [RFC8699] in case all traffic can be assumed to take the same path, or otherwise with a multipath congestion window coupling mechanism as in Multipath TCP [RFC6356]. Another possibility would be to exploit a lower than best-effort congestion control [RFC6297] for repair symbols.

5.1. Fairness and impact on non-coded flows

The coding scheme may hide congestion losses from the congestion controller. There are cases where this can drastically reduce the goodput of non-coded flows. Depending on the congestion control, it may be possible to signal to the congestion control mechanism that there was congestion (loss) even when a packet has been recovered, e.g. using ECN, to reduce the impact on the non-coded flows (see Section 5.2 and [TENTET]).

5.2. Congestion control and recovered symbols

The congestion control may not be aware of the existence of a coding scheme underneath it. The congestion control may behave as if no coding scheme had been introduced. The only way for a coding channel to indicate that symbols have been lost but recovered is to exploit existing signaling that is understood by the congestion control mechanism. An example would be to indicate to a TCP sender that a packet has been received, yet congestion has occurred, by using ECN signaling [TENTET].

5.3. Interactions between congestion control and coding rates

The coding rate can be tuned depending on the number of recovered symbols and the rate at which the sender transmits data. If the coding scheme is not aware of the congestion control implementation, it is hard for the coding scheme to apply the relevant coding rate.

5.4. On useless repair symbols

Useless repair symbols only impact the load on the network without actual gain for the coded flow. Using feedback signaling, FEC mechanisms can measure the ratio between the number of symbols that were actually used and the number of symbols that useless, and adjust the coding rate.

5.5. On partial ordering at FEC level with in-order delivery transport

The transport above the FEC channel may support out-of-order delivery of packets: reordering mechanisms at the receiver may not be necessary. In cases where the transport requires in-order delivery, the FEC channel may need to implement a reordering mechanism. Otherwise, spurious retransmissions may occur at the transport level.

5.6. On partial reliability at FEC level

The transport or application layer above the FEC channel may require partial reliability only. FEC may provide an unnecessary service unless it is aware of the reliability requirements. Partial reliability impacts the type of FEC and type of codec that can be used, such as discussed in Section 2.5.

5.7. FEC not aware of transport multipath

The transport may exploit multiple paths without the FEC channel being aware of it. If FEC is aware that multiple paths are in use, FEC can be applied to all subflows as an aggregate, or to each of the subflows individually. If FEC is not aware that multiple paths are in use, FEC can only be applied to each subflow individually. When FEC is applied to all the flows as an aggregate, the varying characteristics of the individual paths may lead to a risk for the coding rate to be inadequate for the characteristics of the individual paths.

6. Research recommendations and questions

This section provides a short state-of-the art overview of activities related to congestion control and coding. The objective is to identify open research questions and contribute to advice when evaluating coding mechanisms.

6.1. Activities related to congestion control and coding

We map activities related to congestion control and coding with the organization presented in this document:

- * For the FEC above transport case: [RFC8680].
- * For the FEC within transport case:
[I-D.swett-nwcr-g-coding-for-quic], [QUIC-FEC], [RFC5109].
- * For the FEC below transport case: [NCTCP],
[I-D.detchart-nwcr-g-tetrys].

6.2. Open research questions

There is a general trade-off, inherent to the use of coding, between (1) reducing goodput when useless repair symbols are transmitted and (2) helping to recover from transmission and congestion losses.

6.2.1. Parameter derivation

There is a trade-off related to the amount of redundancy to add, as a function of the transport layer protocol and application requirements.

[RFC8095] describes the mechanisms provided by existing IETF protocols such as TCP, SCTP or RTP. [RFC8406] describes the variety of coding techniques. The number of combinations makes the determination of an optimum parameters derivation very complex. This depends on application requirements and deployment context.

Appendix C of [RFC8681] describes how to tune the parameters for target use-case. However, this discussion does not integrate congestion-controlled end points.

Research question 1 : "Is there a way to dynamically adjust the codec characteristics depending on the transmission channel, the transport protocol and application requirements ?"

Research question 2 : "Should we apply specific per-stream FEC mechanisms when multiple streams with different reliability needs are carried out ?"

6.2.2. New signaling methods and fairness

Recovering lost symbols may hide congestion losses from the congestion control. Disambiguate acked packets from rebuilt packets would help the sender adapt its sending rate accordingly. There are opportunities for introducing interaction between congestion control and coding schemes to improve the quality of experience while guaranteeing fairness with other flows.

Some existing solutions already propose to disambiguate acked packets from rebuilt packets [QUIC-FEC]. New signaling methods and FEC-recovery-aware congestion controls could be proposed. This would allow the design of adaptive coding rates.

Research question 3 : "Should we quantify the harm that a coded flow would induce on a non-coded flow ? How can this be reduced while still benefiting from advantages brought by FEC ?"

Research question 4 : "If transport and FEC senders are collocated and close to the client, and FEC is applied only on the last mile, e.g. to ignore losses on a noisy wireless link, would this raise fairness issues ?"

Research question 5 : "Should we propose a generic API to allow dynamic interactions between a transport protocol and a coding scheme ? This should consider existing APIs between application and transport layers."

6.3. Recommendations and advice for evaluating coding mechanisms

Research Recommendation 1: "From a congestion control point-of-view, a recovered packet must be considered as a lost packet. This does not apply to the usage of FEC on a path that is known to be lossy."

Research Recommendation 2: "New research contributions should be mapped following the organization of this document (above, below, in the congestion control) and should consider congestion control aspects when proposing and comparing FEC coding solutions in communication systems."

Research Recommendation 3: "When a research work aims at improving throughput by hiding the packet loss signal from congestion control (e.g., because the path between the sender and receiver is known to consist of a noisy wireless link), the authors should 1) discuss the advantages of using the proposed FEC solution compared to replacing the congestion control by one that ignores a portion of the encountered losses, 2) critically discuss the impact of hiding packet loss from the congestion control mechanism."

7. Acknowledgements

Many thanks to Spencer Dawkins, Dave Oran, Carsten Bormann, Vincent Roca and Marie-Jose Montpetit for their useful comments that helped improve the document.

8. IANA Considerations

This memo includes no request to IANA.

9. Security Considerations

FEC and CC schemes can contribute to DoS attacks. Moreover, the transmission of signaling messages from the client to the server should be protected and reliable otherwise an attacker may compromise FEC rate adaptation. Indeed, an attacker could either modify the values indicated by the client or drop signaling messages.

In case of FEC below the transport, the aggregate rate of source and repair packets may exceed the rate at which a congestion control mechanism allows an application to send. This could result in an application obtaining more than its fair share of the network capacity.

10. Informative References

[BEYONDJAIN]

Ware (et al.), R., "Beyond Jain's Fairness Index: Setting the Bar For The Deployment of Congestion Control Algorithms", HotNets '19 10.1145/3365609.3365855, 2019.

[CTCP]

Kim (et al.), M., "Network Coded TCP (CTCP)", arXiv 1212.2291v3, 2013.

[I-D.briscoe-tsvarea-fair]

Briscoe, B., "Flow Rate Fairness: Dismantling a Religion", Work in Progress, Internet-Draft, draft-briscoe-tsvarea-fair-02, 11 July 2007, <<https://www.ietf.org/archive/id/draft-briscoe-tsvarea-fair-02.txt>>.

[I-D.detchart-nwcr-g-tetrys]

Detchart, J., Lochin, E., Lacan, J., and V. Roca, "Tetrys, an On-the-Fly Network Coding protocol", Work in Progress, Internet-Draft, draft-detchart-nwcr-g-tetrys-08, 17 October 2021, <<https://www.ietf.org/archive/id/draft-detchart-nwcr-g-tetrys-08.txt>>.

[I-D.ietf-quic-datagram]

Pauly, T., Kinnear, E., and D. Schinazi, "An Unreliable Datagram Extension to QUIC", Work in Progress, Internet-Draft, draft-ietf-quic-datagram-10, 4 February 2022, <<https://www.ietf.org/archive/id/draft-ietf-quic-datagram-10.txt>>.

[I-D.singh-rmcat-adaptive-fec]

Singh, V., Nagy, M., Ott, J., and L. Eggert, "Congestion Control Using FEC for Conversational Media", Work in Progress, Internet-Draft, draft-singh-rmcat-adaptive-fec-03, 20 March 2016, <<https://www.ietf.org/archive/id/draft-singh-rmcat-adaptive-fec-03.txt>>.

- [I-D.swett-nwcr-g-coding-for-quic]
Swett, I., Montpetit, M., Roca, V., and F. Michel, "Coding for QUIC", Work in Progress, Internet-Draft, draft-swett-nwcr-g-coding-for-quic-04, 9 March 2020, <<https://www.ietf.org/archive/id/draft-swett-nwcr-g-coding-for-quic-04.txt>>.
- [NCTCP] Sundararajan (et al.), J., "Network Coding Meets TCP: Theory and Implementation", IEEE INFOCOM 10.1109/JPROC.2010.2093850, 2009.
- [QUIC-FEC] Michel (et al.), F., "QUIC-FEC: Bringing the benefits of Forward Erasure Correction to QUIC", IFIP Networking 10.23919/IFIPNetworking.2019.8816838, 2019.
- [RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <<https://www.rfc-editor.org/info/rfc3168>>.
- [RFC3758] Stewart, R., Ramalho, M., Xie, Q., Tuexen, M., and P. Conrad, "Stream Control Transmission Protocol (SCTP) Partial Reliability Extension", RFC 3758, DOI 10.17487/RFC3758, May 2004, <<https://www.rfc-editor.org/info/rfc3758>>.
- [RFC4340] Kohler, E., Handley, M., and S. Floyd, "Datagram Congestion Control Protocol (DCCP)", RFC 4340, DOI 10.17487/RFC4340, March 2006, <<https://www.rfc-editor.org/info/rfc4340>>.
- [RFC5109] Li, A., Ed., "RTP Payload Format for Generic Forward Error Correction", RFC 5109, DOI 10.17487/RFC5109, December 2007, <<https://www.rfc-editor.org/info/rfc5109>>.
- [RFC5681] Allman, M., Paxson, V., and E. Blanton, "TCP Congestion Control", RFC 5681, DOI 10.17487/RFC5681, September 2009, <<https://www.rfc-editor.org/info/rfc5681>>.
- [RFC6297] Welzl, M. and D. Ros, "A Survey of Lower-than-Best-Effort Transport Protocols", RFC 6297, DOI 10.17487/RFC6297, June 2011, <<https://www.rfc-editor.org/info/rfc6297>>.
- [RFC6356] Raiciu, C., Handley, M., and D. Wischik, "Coupled Congestion Control for Multipath Transport Protocols", RFC 6356, DOI 10.17487/RFC6356, October 2011, <<https://www.rfc-editor.org/info/rfc6356>>.

- [RFC8095] Fairhurst, G., Ed., Trammell, B., Ed., and M. Kuehlewind, Ed., "Services Provided by IETF Transport Protocols and Congestion Control Mechanisms", RFC 8095, DOI 10.17487/RFC8095, March 2017, <<https://www.rfc-editor.org/info/rfc8095>>.
- [RFC8406] Adamson, B., Adjih, C., Bilbao, J., Firoiu, V., Fitzek, F., Ghanem, S., Lochin, E., Masucci, A., Montpetit, M-J., Pedersen, M., Peralta, G., Roca, V., Ed., Saxena, P., and S. Sivakumar, "Taxonomy of Coding Techniques for Efficient Network Communications", RFC 8406, DOI 10.17487/RFC8406, June 2018, <<https://www.rfc-editor.org/info/rfc8406>>.
- [RFC8680] Roca, V. and A. Begen, "Forward Error Correction (FEC) Framework Extension to Sliding Window Codes", RFC 8680, DOI 10.17487/RFC8680, January 2020, <<https://www.rfc-editor.org/info/rfc8680>>.
- [RFC8681] Roca, V. and B. Teibi, "Sliding Window Random Linear Code (RLC) Forward Erasure Correction (FEC) Schemes for FECFRAME", RFC 8681, DOI 10.17487/RFC8681, January 2020, <<https://www.rfc-editor.org/info/rfc8681>>.
- [RFC8699] Islam, S., Welzl, M., and S. Gjessing, "Coupled Congestion Control for RTP Media", RFC 8699, DOI 10.17487/RFC8699, January 2020, <<https://www.rfc-editor.org/info/rfc8699>>.
- [TENTET] Lochin, E., "On the joint use of TCP and Network Coding", NWCRG session IETF 100, 2017.

Authors' Addresses

Nicolas Kuhn
CNES
Email: nicolas.kuhn.ietf@gmail.com

Emmanuel Lochin
ENAC
Email: emmanuel.lochin@enac.fr

Francois Michel
UCLouvain
Email: francois.michel@uclouvain.be

Michael Welzl
University of Oslo
Email: michawe@ifi.uio.no