On Elias-Fano for Rank Queries in FM-Indexes

Danyang Ma, Simon J. Puglisi, Rajeev Raman, and Bella Zhukova

DCC 2021

Definitions: BWT

$T = \text{mississippi\$}$ $\Sigma = \{\$, i, m, p, s\}$

m	i	s	s	i	s	s	i	p	p	i	$
i	s	s	i	s	s	i	p	p	i	$	m
s	s	i	s	s	i	p	p	i	$	m	i
s	i	s	s	i	p	p	i	$	m	i	s
i	s	s	i	p	p	i	$	m	i	s	s
s	s	i	p	p	i	$	m	i	s	s	i
s	i	p	p	i	$	m	i	s	s	i	s
i	p	p	i	$	m	i	s	s	i	s	s
p	p	i	$	m	i	s	s	i	s	s	i
p	i	$	m	i	s	s	i	s	s	i	p
i	$	m	i	s	s	i	s	s	i	p	p
$	m	i	s	s	i	s	s	i	p	p	i

Definitions: BWT

$T = \text{mississippi\$}$ $\Sigma = \{\$, i, m, p, s\}$

5	m	i	s	s	i	s	s	i	p	p	i	$
4	i	s	s	i	s	s	i	p	p	i	$	m
11	s	s	i	s	s	i	p	p	i	$	m	i
9	s	i	s	s	i	p	p	i	$	m	i	s
3	i	s	s	i	p	p	i	$	m	i	s	s
10	s	s	i	p	p	i	$	m	i	s	s	i
8	s	i	p	p	i	$	m	i	s	s	i	s
2	i	p	p	i	$	m	i	s	s	i	s	s
7	p	p	i	$	m	i	s	s	i	s	s	i
6	p	i	$	m	i	s	s	i	s	s	i	p
1	i	$	m	i	s	s	i	s	s	i	p	p
0	$	m	i	s	s	i	s	s	i	p	p	i

Definitions: BWT

$T = \text{mississippi\$}$ $\Sigma = \{\$, i, m, p, s\}$

0	\$	m	i	s	s	i	s	s	i	p	p	i
1	i	\$	m	i	s	s	i	s	s	i	p	p
2	i	p	p	i	\$	m	i	s	s	i	s	s
3	i	s	s	i	p	p	i	\$	m	i	s	s
4	i	s	s	i	s	s	i	p	p	i	\$	m
5	m	i	s	s	i	s	s	i	p	p	i	\$
6	p	i	\$	m	i	s	s	i	s	s	i	p
7	p	p	i	\$	m	i	s	s	i	s	s	i
8	s	i	p	p	i	\$	m	i	s	s	i	s
9	s	i	s	s	i	p	p	i	\$	m	i	s
10	s	s	i	p	p	i	\$	m	i	s	s	i
11	s	s	i	s	s	i	p	p	i	\$	m	i

Definitions: BWT

$T = \text{mississippi\$}$ $\Sigma = \{\$, i, m, p, s\}$

	F										
0	$	m	i	s	s	i	s	s	i	p	p	i
1	i	$	m	i	s	s	i	s	s	i	p	p
2	i	p	p	i	$	m	i	s	s	i	s	s
3	i	s	s	i	p	p	i	$	m	i	s	s
4	i	s	s	i	s	s	i	p	p	i	$	m
5	m	i	s	s	i	s	s	i	p	p	i	$
6	p	i	$	m	i	s	s	i	s	s	i	p
7	p	p	i	$	m	i	s	s	i	s	s	i
8	s	i	p	p	i	$	m	i	s	s	i	s
9	s	i	s	s	i	p	p	i	$	m	i	s
10	s	s	i	p	p	i	$	m	i	s	s	i
11	s	s	i	s	s	i	p	p	i	$	m	i

Definitions: BWT

$T = \text{mississippi\$}$ $\Sigma = \{\$, i, m, p, s\}$

	F										L, BWT
0	\$	m	i	s	s	i	s	s	i	p	p	i
1	i	\$	m	i	s	s	i	s	s	i	p	p
2	i	p	p	i	\$	m	i	s	s	i	s	s
3	i	s	s	i	p	p	i	\$	m	i	s	s
4	i	s	s	i	s	s	i	p	p	i	\$	m
5	m	i	s	s	i	s	s	i	p	p	i	\$
6	p	i	\$	m	i	s	s	i	s	s	i	p
7	p	p	i	\$	m	i	s	s	i	s	s	i
8	s	i	p	p	i	\$	m	i	s	s	i	s
9	s	i	s	s	i	p	p	i	\$	m	i	s
10	s	s	i	p	p	i	\$	m	i	s	s	i
11	s	s	i	s	s	i	p	p	i	\$	m	i

Problem (Counting Queries)

For a pattern P and text T, *count(P)* returns the number of occurrences of P in T

Definitions: *rank* and *select*

$rank_X(i,\ c)$ on string X

is the number of occurrences of symbol c in prefix $X[0,\ i\text{-}1]$

$select_X(i,\ c)$ on string X

is the position of the i-th occurrence of symbol c

Definitions: BWT

$T = \text{mississippi\$}$

$\Sigma = \{\$, i, m, p, s\}$

	F	L, BWT
0	\$	i
1	i	p
2	i	s
3	i	s
4	i	m
5	m	\$
6	p	p
7	p	i
8	s	s
9	s	s
10	s	i
11	s	i

	0	1	2	3	4	5	6	7	8	9	10	11
\$:	0	0	0	0	0	1	0	0	0	0	0	0
i:	1	0	0	0	0	0	0	1	0	0	1	1
m:	0	0	0	0	1	0	0	0	0	0	0	0
p:	0	1	0	0	0	0	1	0	0	0	0	0
s:	0	0	1	1	0	0	0	0	1	1	0	0

Definitions: BWT

$T = \text{mississippi\$}$ $\Sigma = \{\$, i, m, p, s\}$ $P = \text{isi}$

	F	L, BWT
0	\$	i
1	i	p
2	i	s
3	i	s
4	i	m
5	m	\$
6	p	p
7	p	i
8	s	s
9	s	s
10	s	i
11	s	i

	0	1	2	3	4	5	6	7	8	9	10	11
\$:	0	0	0	0	0	1	0	0	0	0	0	0
i:	1	0	0	0	0	0	0	1	0	0	1	1
m:	0	0	0	0	1	0	0	0	0	0	0	0
p:	0	1	0	0	0	0	1	0	0	0	0	0
s:	0	0	1	1	0	0	0	0	1	1	0	0

Definitions: BWT

$T =$ mississippi\$ $\quad \Sigma = \{\$, i, m, p, s\}$ $\quad P =$ isi

	F	L, BWT
0	\$	i
1	i	p
2	i	s
3	i	s
4	i	m
5	m	\$
6	p	p
7	p	i
8	s	s
9	s	s
10	s	i
11	s	i

	0	1	2	3	4	5	6	7	8	9	10	11
\$:	0	0	0	0	0	1	0	0	0	0	0	0
i:	1	0	0	0	0	0	0	1	0	0	1	1
m:	0	0	0	0	1	0	0	0	0	0	0	0
p:	0	1	0	0	0	0	1	0	0	0	0	0
s:	0	0	1	1	0	0	0	0	1	1	0	0

Definitions: BWT

$T =$ mississippi\$ $\quad \Sigma = \{\$, i, m, p, s\}$ $\quad P =$ isi

	F	L, BWT
0	\$	i
1	i	p
2	i	s
3	i	s
4	i	m
5	m	\$
6	p	p
7	p	i
8	s	s
9	s	s
10	s	i
11	s	i

	0	1	2	3	4	5	6	7	8	9	10	11
\$:	0	0	0	0	0	1	0	0	0	0	0	0
i:	1	0	0	0	0	0	0	1	0	0	1	1
m:	0	0	0	0	1	0	0	0	0	0	0	0
p:	0	1	0	0	0	0	1	0	0	0	0	0
s:	0	0	1	1	0	0	0	0	1	1	0	0

Definitions: BWT

$T =$ mississippi\$ $\quad \Sigma = \{\$, i, m, p, s\}$ $\quad P =$ isi

	F	L, BWT
0	\$	i
1	i	p
2	i	s
3	i	s
4	i	m
5	m	\$
6	p	p
7	p	i
8	s	s
9	s	s
10	s	i
11	s	i

	0	1	2	3	4	5	6	7	8	9	10	11
\$:	0	0	0	0	0	1	0	0	0	0	0	0
i:	1	0	0	0	0	0	0	1	0	0	1	1
m:	0	0	0	0	1	0	0	0	0	0	0	0
p:	0	1	0	0	0	0	1	0	0	0	0	0
s:	0	0	1	1	0	0	0	0	1	1	0	0

Definitions: BWT

$T =$ mississippi\$ $\quad \Sigma = \{\$, i, m, p, s\}$ $\quad P =$ isi

	F	L, BWT
0	\$	i
1	i	p
2	i	s
3	i	s
4	i	m
5	m	\$
6	p	p
7	p	i
8	s	s
9	s	s
10	s	i
11	s	i

	0	1	2	3	4	5	6	7	8	9	10	11
\$:	0	0	0	0	0	1	0	0	0	0	0	0
i:	1	0	0	0	0	0	0	1	0	0	1	1
m:	0	0	0	0	1	0	0	0	0	0	0	0
p:	0	1	0	0	0	0	1	0	0	0	0	0
s:	0	0	1	1	0	0	0	0	1	1	0	0

Definitions: BWT

$T =$ mississippi\$ $\quad \Sigma = \{\$, i, m, p, s\}$ $\quad P =$ isi

	F	L, BWT
0	\$	i
1	i	p
2	i	s
3	i	s
4	i	m
5	m	\$
6	p	p
7	p	i
8	s	s
9	s	s
10	s	i
11	s	i

	0	1	2	3	4	5	6	7	8	9	10	11
\$:	0	0	0	0	0	1	0	0	0	0	0	0
i:	1	0	0	0	0	0	0	1	0	0	1	1
m:	0	0	0	0	1	0	0	0	0	0	0	0
p:	0	1	0	0	0	0	1	0	0	0	0	0
s:	0	0	1	1	0	0	0	0	1	1	0	0

Previous Work

- Many authors at this bit-vector representation on BWT (Mäkinen, Navarro, Huo, Sirén, Grabowski,...)

- We are re-examined the use of Elias-Fano for representing bit-vectors

Our Contribution

- Optimizations on Elias-Fano that quicken rank query
  - may be useful in other problems/contexts, not only for counting queries

- Two Elias-Fano-like variants independent of select

Elias-Fano

	0	1	2	3	4	5	6	7	8	9	10	11
\$:	0	0	0	0	0	1	0	0	0	0	0	0
i:	1	0	0	0	0	0	0	1	0	0	1	1
m:	0	0	0	0	1	0	0	0	0	0	0	0
p:	0	1	0	0	0	0	1	0	0	0	0	0
s:	0	0	1	1	0	0	0	0	1	1	0	0

Elias-Fano

	0	1	2	3	4	5	6	7	8	9	10	11
\$:	0	0	0	0	0	1	0	0	0	0	0	0
i:	1	0	0	0	0	0	0	1	0	0	1	1
m:	0	0	0	0	1	0	0	0	0	0	0	0
p:	0	1	0	0	0	0	1	0	0	0	0	0
s:	0	0	1	1	0	0	0	0	1	1	0	0

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$

Elias-Fano

$$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$$

$$5 \quad 12 \quad 19 \quad 22 \quad 23 \quad 28 \quad 37 \quad 42 \quad 50 \quad 51 \quad 56 \quad 57$$

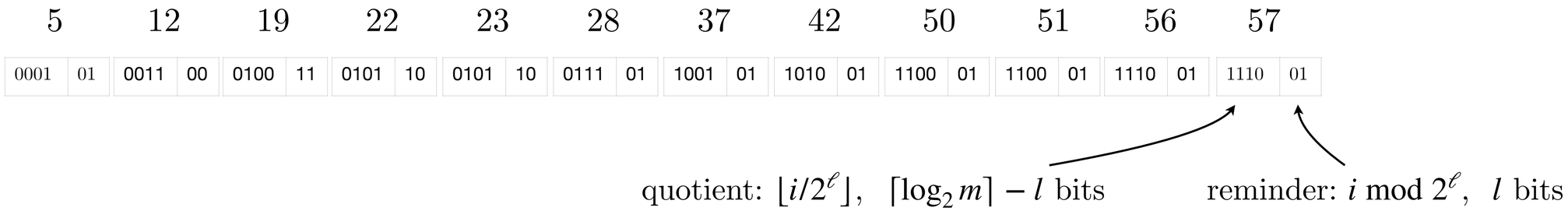
Elias-Fano

$m = 57,\ n = 12,\ l = 2$

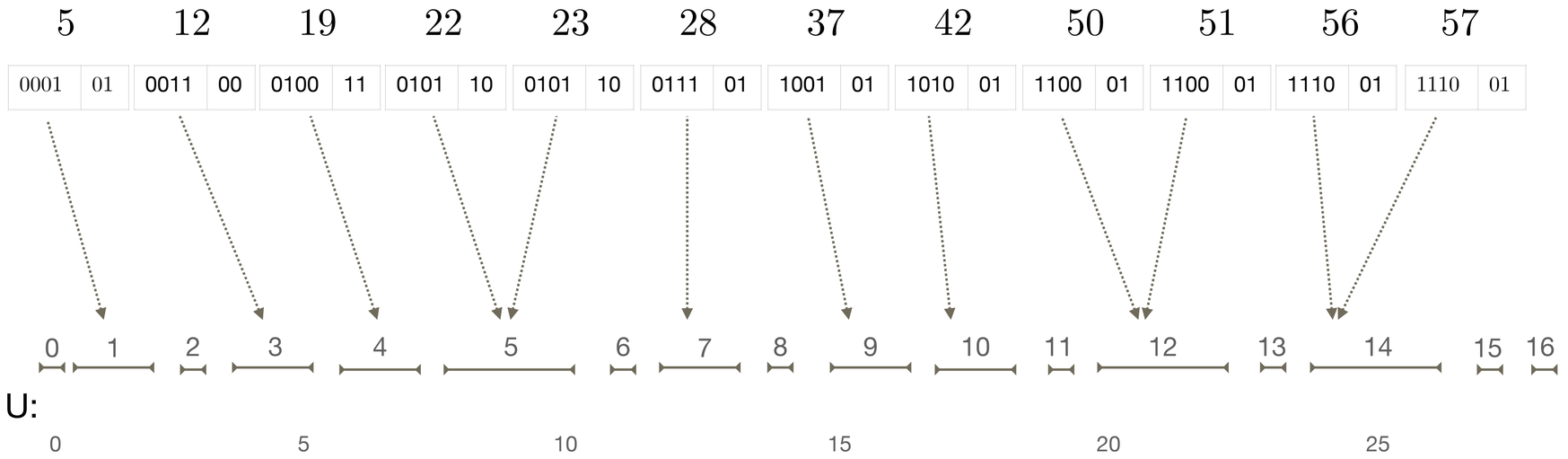
5 12 19 22 23 28 37 42 50 51 56 57

Elias-Fano

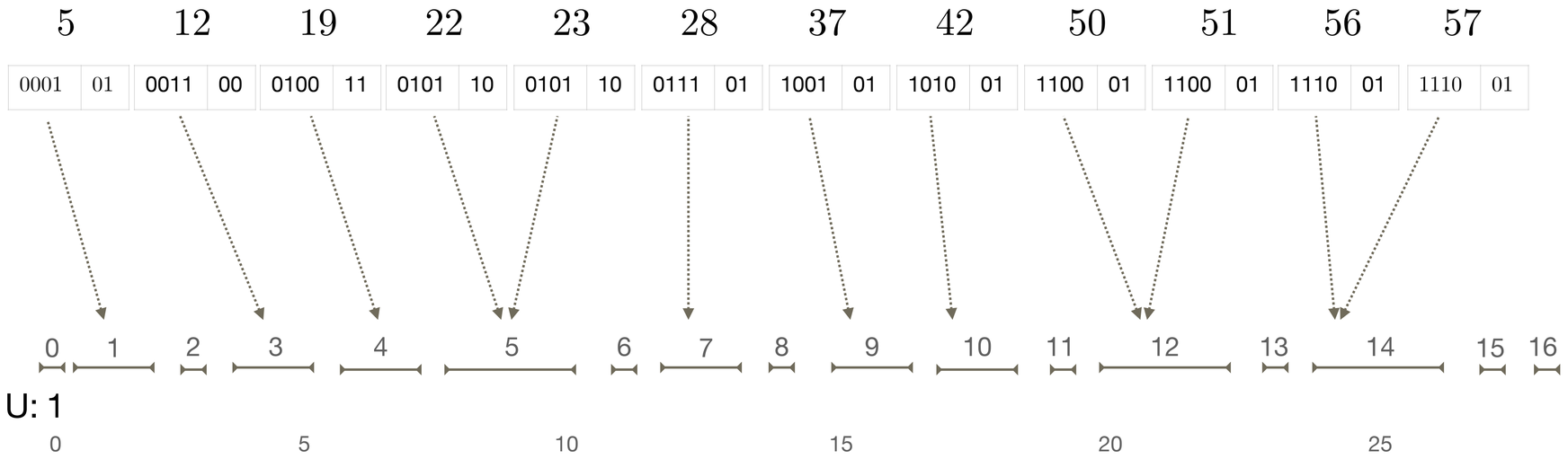
$m = 57,\ n = 12,\ l = 2$



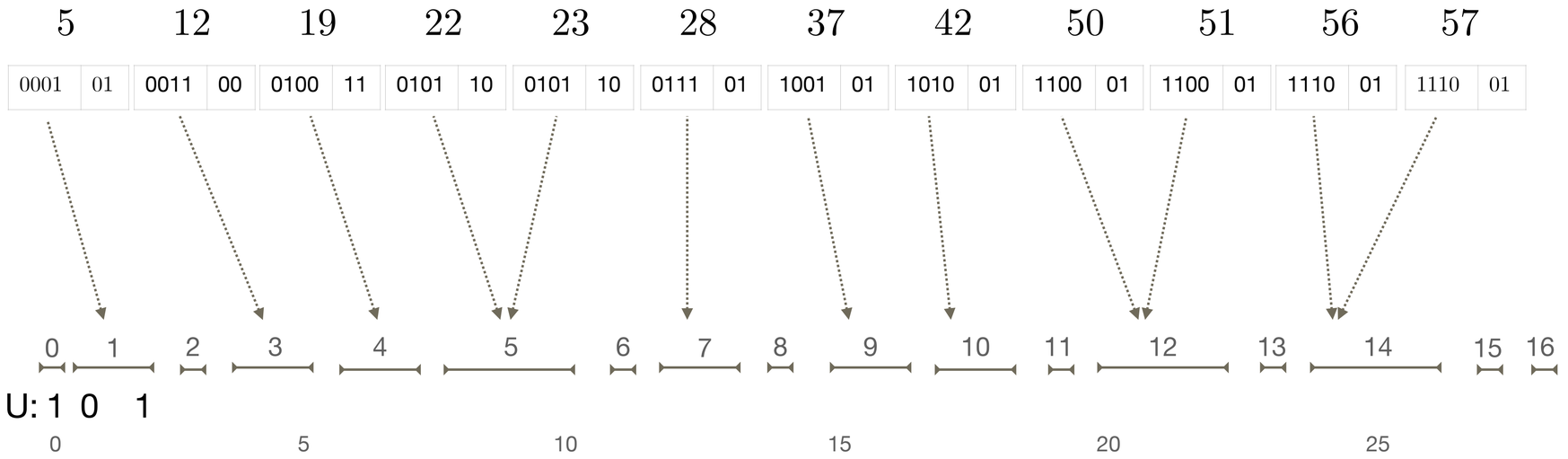
Elias-Fano


5 12 19 22 23 28 37 42 50 51 56 57
0001 01
0011 00
0100 11
0101 10
0101 10
0111 01
1001 01
1010 01
1100 01
1100 01
1110 01
1110 01
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
U:
0 5 10 15 20 25

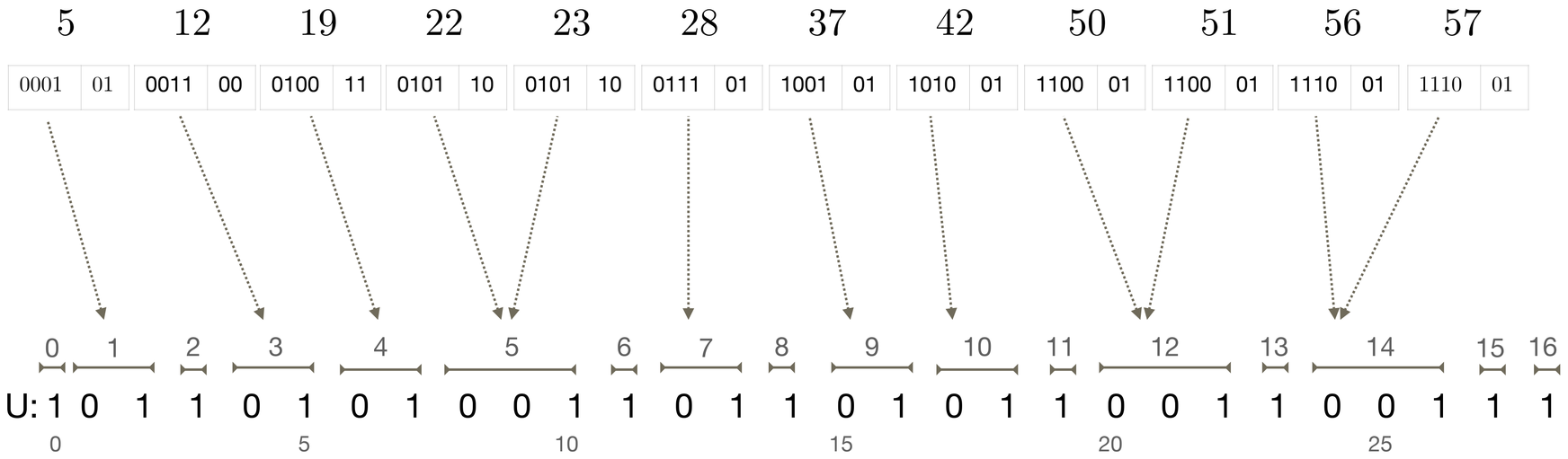
Elias-Fano


5 12 19 22 23 28 37 42 50 51 56 57
0001 01
0011 00
0100 11
0101 10
0101 10
0111 01
1001 01
1010 01
1100 01
1100 01
1110 01
1110 01
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
U: 1
0 5 10 15 20 25

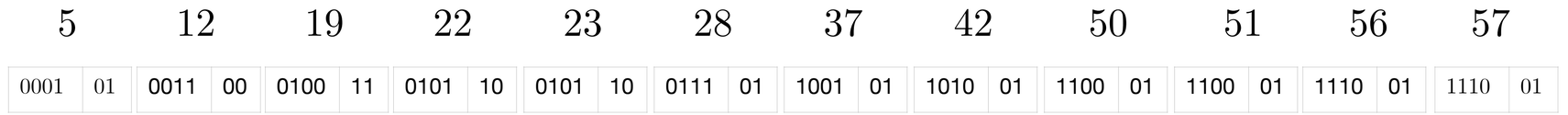
Elias-Fano

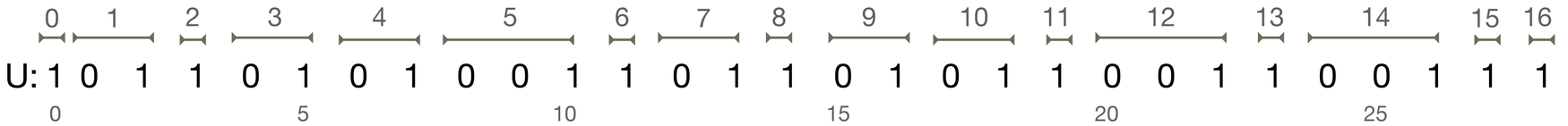

5 12 19 22 23 28 37 42 50 51 56 57
0001 01
0011 00
0100 11
0101 10
0101 10
0111 01
1001 01
1010 01
1100 01
1100 01
1110 01
1110 01
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
U: 1 0 1
0 5 10 15 20 25

Elias-Fano


5 12 19 22 23 28 37 42 50 51 56 57
0001 01 0011 00 0100 11 0101 10 0101 10 0111 01 1001 01 1010 01 1100 01 1100 01 1110 01 1110 01
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
U: 1 0 1 1 0 1 0 1 0 0 1 1 0 1 1 0 1 0 1 1 0 0 1 1 0 0 1 1 1
0 5 10 15 20 25

Elias-Fano

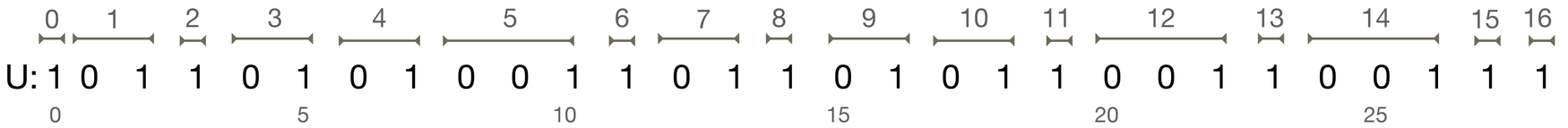

5 12 19 22 23 28 37 42 50 51 56 57
0001 01 0011 00 0100 11 0101 10 0101 10 0111 01 1001 01 1010 01 1100 01 1100 01 1110 01 1110 01


0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
U: 1 0 1 1 0 1 0 1 0 0 1 1 0 1 1 0 1 0 1 1 0 0 1 1 0 0 1 1 1
0 5 10 15 20 25
L: 01 00 11 10 11 00 01 10 10 11 00 01

Elias-Fano

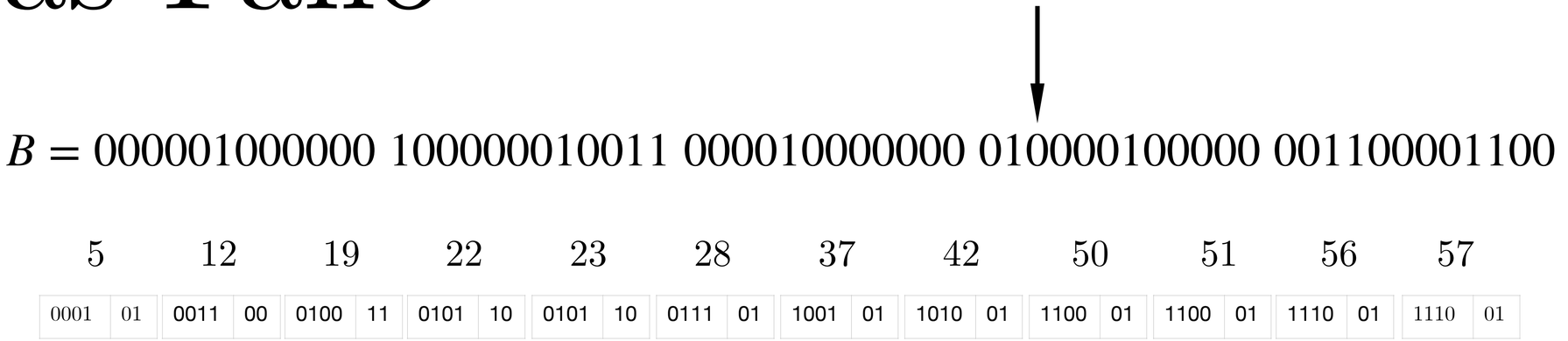
$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$

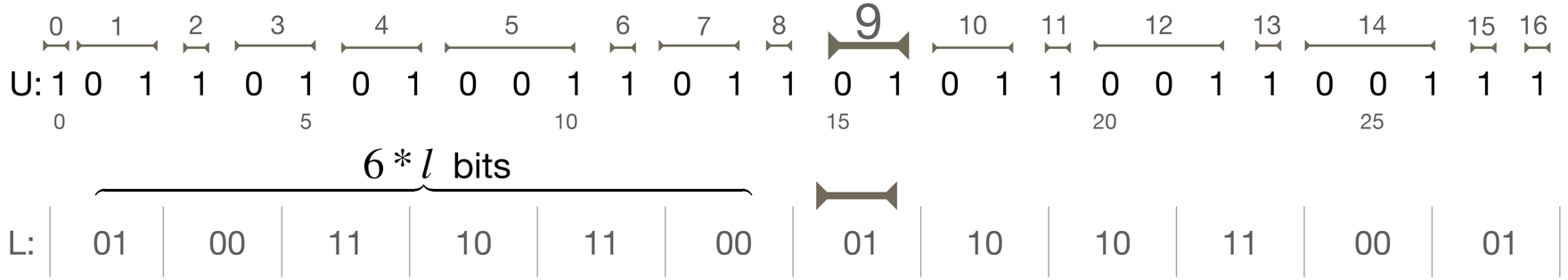
5	12	19	22	23	28	37	42	50	51	56	57
0001 01	0011 00	0100 11	0101 10	0101 10	0111 01	1001 01	1010 01	1100 01	1100 01	1110 01	1110 01



$38 = 1001\,10,\ 1001 \implies 9$

Elias-Fano





$38 = 1001\,10,\ 1001 \implies 9$

$\text{select}_U(9, 1) = 14,\ \text{select}_U(9 + 1, 1) = 16$

$15 - 9 = 6, \qquad 17 - 10 = 7$

Experiment

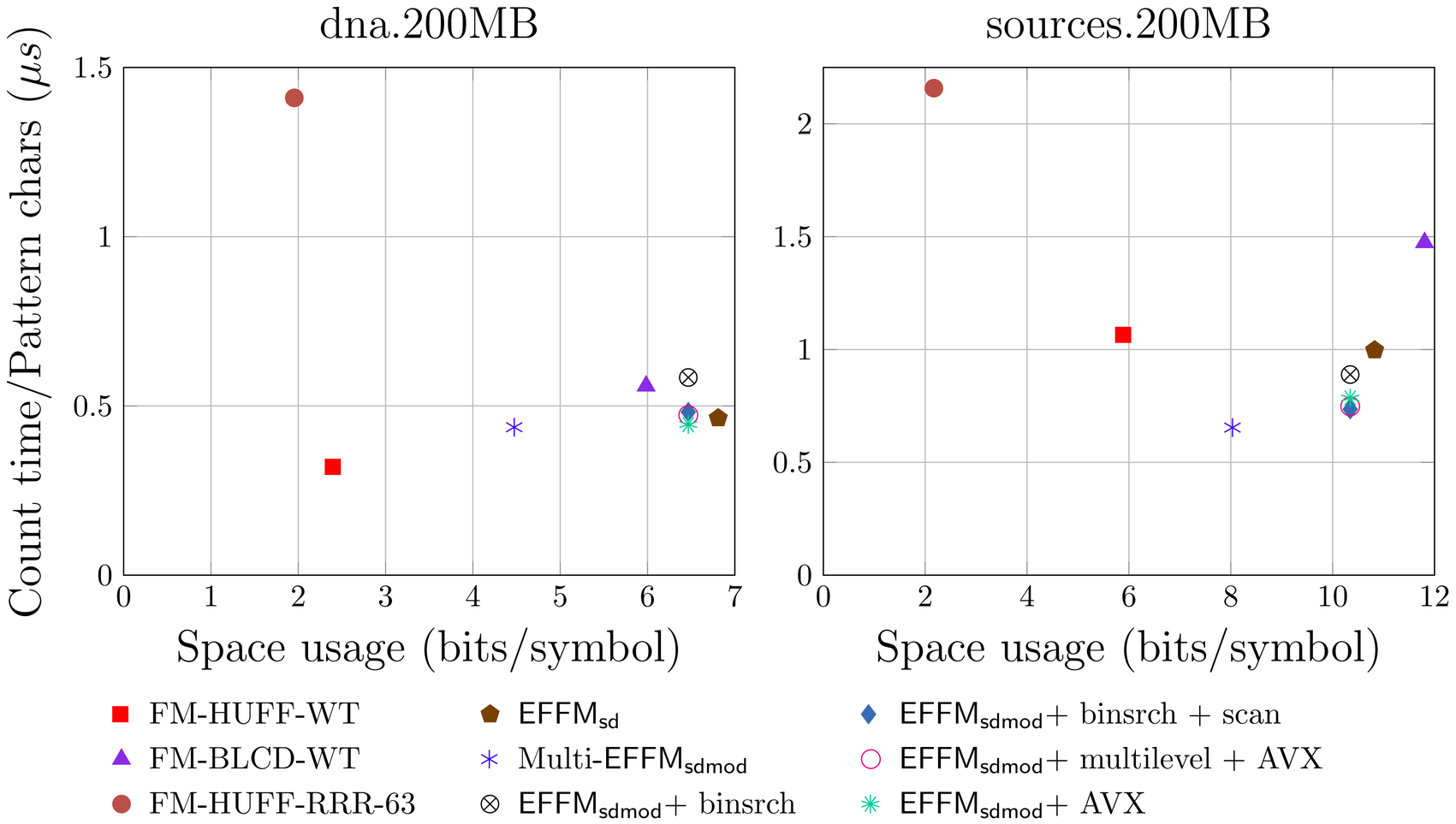
Datasets, 200MB, from the Pizza&Chili corpus:

- DNA
- english
- sources
- XML

Search queries:

- 50 000 patterns
- length = 20

Elias-Fano


dna.200MB
sources.200MB
Count time/Pattern chars (μs)
Space usage (bits/symbol)
Space usage (bits/symbol)
FM-HUFF-WT
FM-BLCD-WT
FM-HUFF-RRR-63
EFFM$_{sd}$
Multi-EFFM$_{sdmod}$
EFFM$_{sdmod}$+ binsrch
EFFM$_{sdmod}$+ binsrch + scan
EFFM$_{sdmod}$+ multilevel + AVX
EFFM$_{sdmod}$+ AVX

New Approaches

- $\mathrm{EFFM}_{\mathrm{zs}}$
  - Zero-Suppressed
- $\mathrm{EFFM}_{\mathrm{pu}}$
  - Partitioned-Upper

New Approaches: EFFM$_{\text{ZS}}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$

let $l = 2$

U	bit vector of length $(n\sigma)/2^{\ell}$, and $U[i] = 1$ iff the i-th bucket is not all zeros
L	$2^{\ell} \cdot$ number of non-zero buckets

New Approaches: $\mathrm{EFFM}_{\mathrm{zs}}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$

let $l = 2$

New Approaches: EFFM_{zs}

$B = 0000\,01000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$

let $l = 2$

New Approaches: EFFM$_{\text{ZS}}$

$B =$ 0000 01000000 100000010011 000010000000 010000100000 001100001100 $\quad \Sigma = \{\$, i, m, p, s\}$

let $l = 2$

New Approaches: EFFM_{zs}

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$

let $l = 2$

New Approaches: EFFM_{zs}

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$

let $l = 2$

New Approaches: EFFM$_{\text{ZS}}$

$B = 000001000000100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$

let $l = 2$

New Approaches: EFFM$_{\text{ZS}}$

$B = 000001000000\ 1000\ 00010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$

let $l = 2$

New Approaches: EFFM$_{\text{ZS}}$

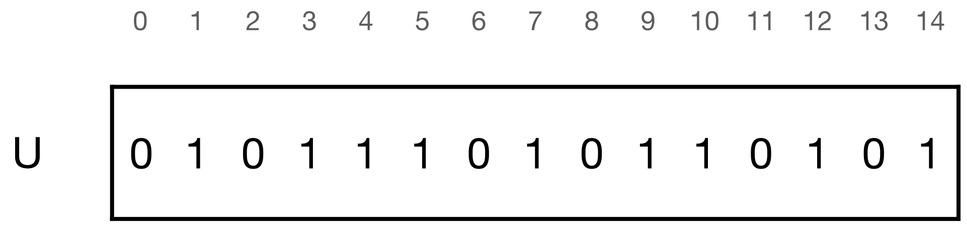
$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$

let $l = 2$

New Approaches: EFFM$_{\text{ZS}}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$

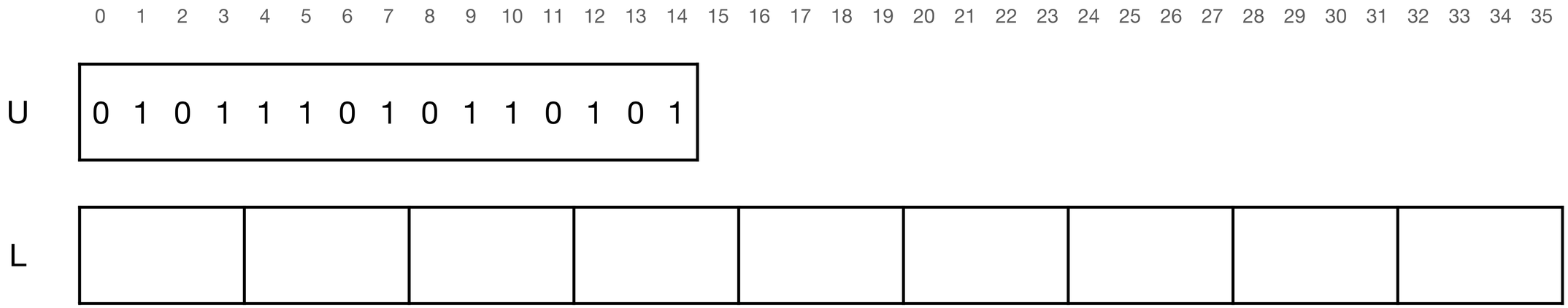
let $l = 2$



New Approaches: EFFM$_{\text{ZS}}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$

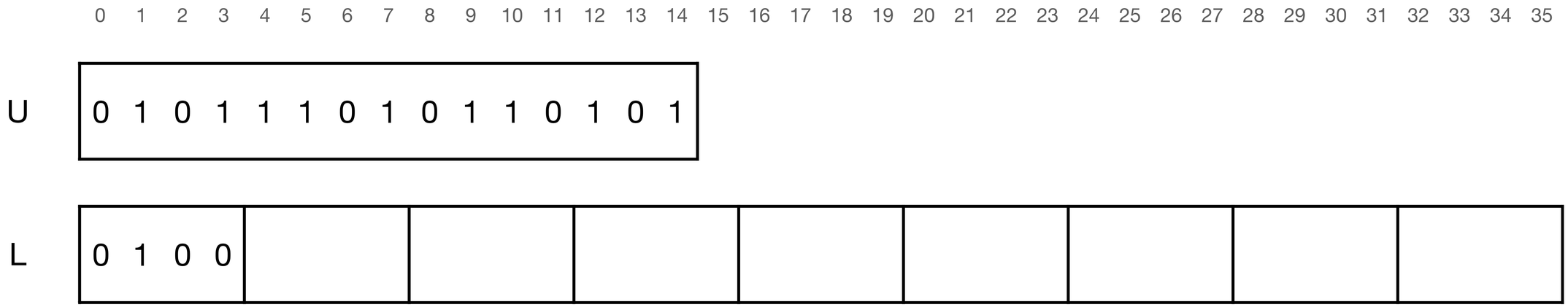
let $l = 2$



New Approaches: EFFM$_{zs}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\Sigma = \{\$, i, m, p, s\}$

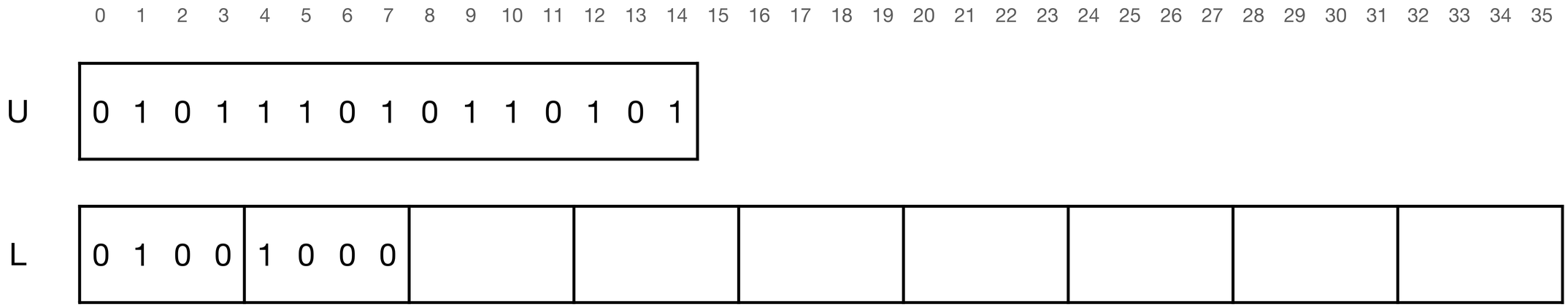
let $l = 2$



New Approaches: EFFM$_{zs}$

$B = 000001000000\ 1000\ 00010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$

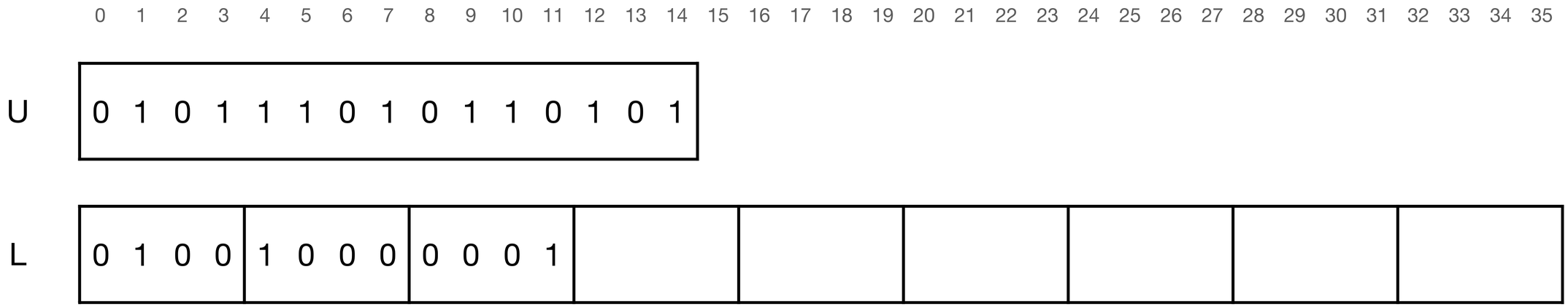
let $l = 2$



New Approaches: EFFM$_{zs}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\Sigma = \{\$, i, m, p, s\}$

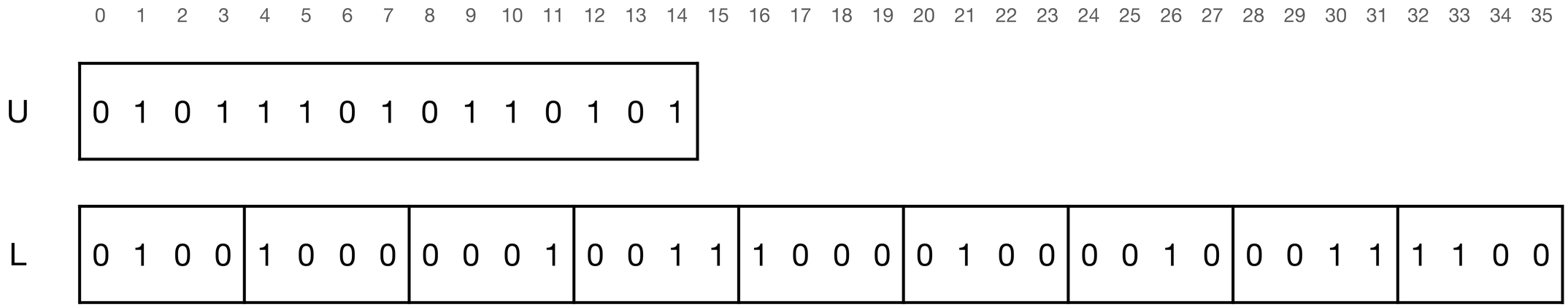
let $l = 2$



New Approaches: $\mathrm{EFFM}_{\mathrm{zs}}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$

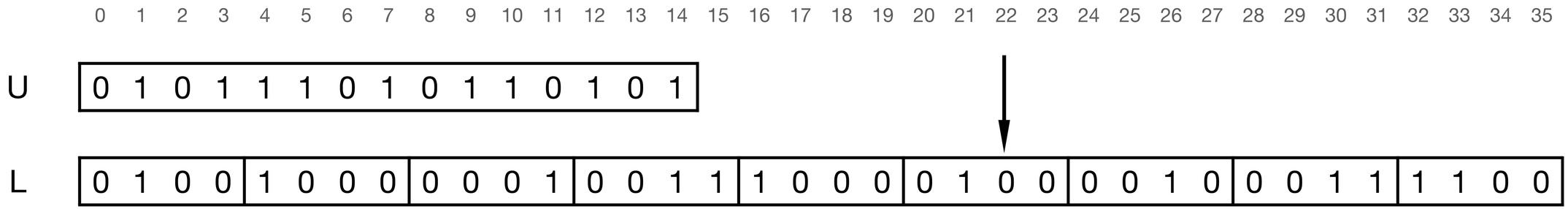
let $l = 2$



New Approaches: EFFM$_{\text{ZS}}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100 \qquad \Sigma = \{\$, i, m, p, s\}$

let $l = 2$



$\text{rank}_B(i, 1)$: $\quad j = \text{rank}_U(\lfloor i/2^\ell \rfloor, 1)$

if $U[\lfloor i/2^\ell \rfloor] = 0 \quad \text{rank}_B(i, 1) = \text{rank}_L(j \cdot 2^\ell, 1)$

if $U[\lfloor i/2^\ell \rfloor] = 1 \quad \text{rank}_B(i, 1) = \text{rank}_L(j \cdot 2^\ell + i \bmod 2^\ell, 1)$

New Approaches: $\mathrm{EFFM}_{\mathrm{pu}}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$
$l = 2$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
U_{nz}	0	1	0	1	1	1	0	1	0	1	1	0	1	0	1
U_{sz}															

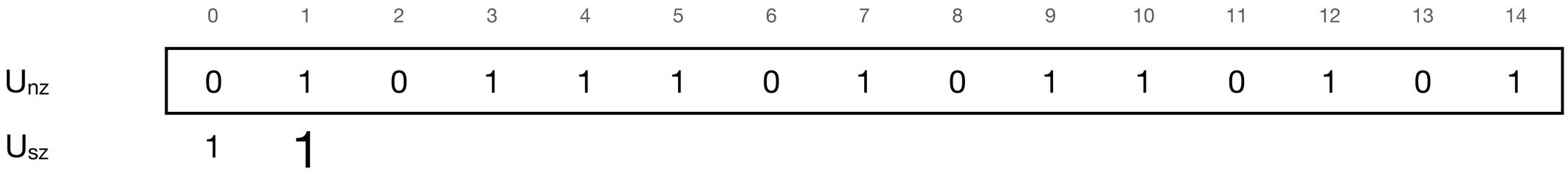
New Approaches: $\mathrm{EFFM}_{\mathrm{pu}}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\Sigma = \{\$, i, m, p, s\}$
$l = 2$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
U_{nz}	0	1	0	1	1	1	0	1	0	1	1	0	1	0	1
U_{sz}	1														

New Approaches: $\mathrm{EFFM}_{\mathrm{pu}}$

$B = 000001000000\ 1000\ 00010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$
$l = 2$



New Approaches: $\mathrm{EFFM}_{\mathrm{pu}}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$
$l = 2$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
U_{nz}	0	1	0	1	1	1	0	1	0	1	1	0	1	0	1
U_{sz}	1	1	1												

New Approaches: $\mathrm{EFFM}_{\mathrm{pu}}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$
$l = 2$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
U_{nz}	0	1	0	1	1	1	0	1	0	1	1	0	1	0	1
U_{sz}	1	1	1	0	1										

New Approaches: EFFM$_{\text{pu}}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\Sigma = \{\$, i, m, p, s\}$
$l = 2$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
U_{nz}	0	1	0	1	1	1	0	1	0	1	1	0	1	0	1
U_{sz}	1	1	1	0	1	1									

New Approaches: EFFM_{pu}

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100 \quad \Sigma = \{\$, i, m, p, s\}$
$l = 2$

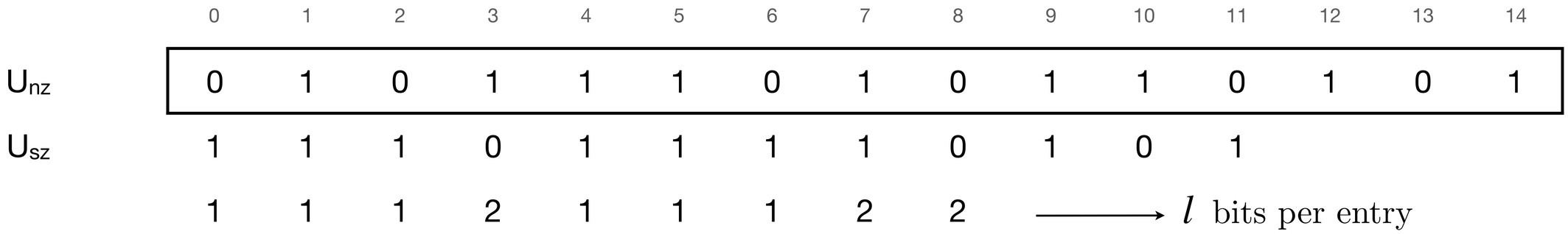
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
U_{nz}	0	1	0	1	1	1	0	1	0	1	1	0	1	0	1
U_{sz}	1	1	1	0	1	1	1	1	0	1	0	1			

L as in sdsl

$$\text{select}_U(\lfloor i/2^\ell \rfloor, 1) = \text{select}_{U_{sz}}(j, 1), \text{ where } j = \text{rank}_{U_{nz}}(\lfloor i/2^\ell \rfloor, 1)$$

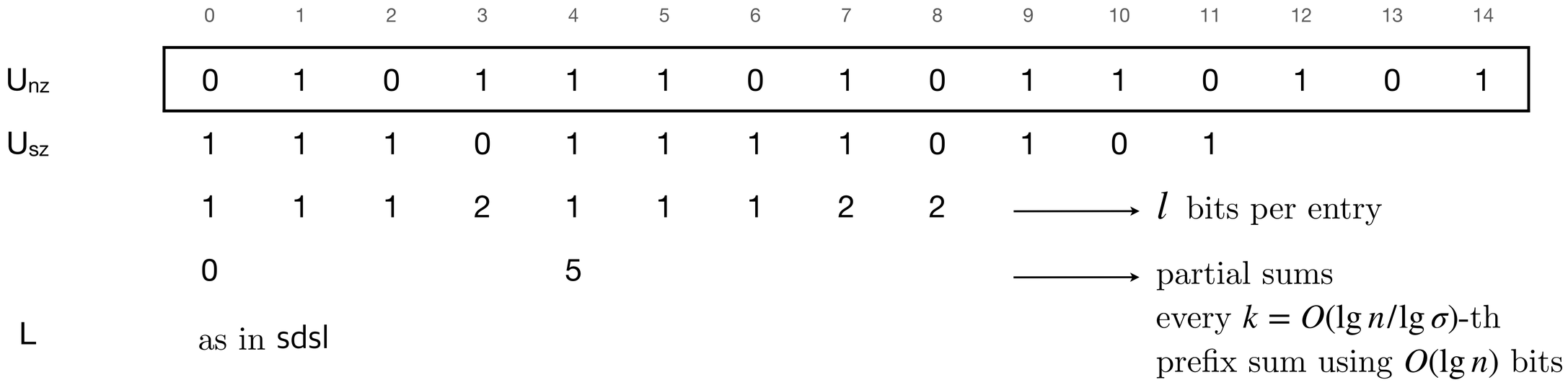
New Approaches: EFFM_{pu}

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$
$l = 2$

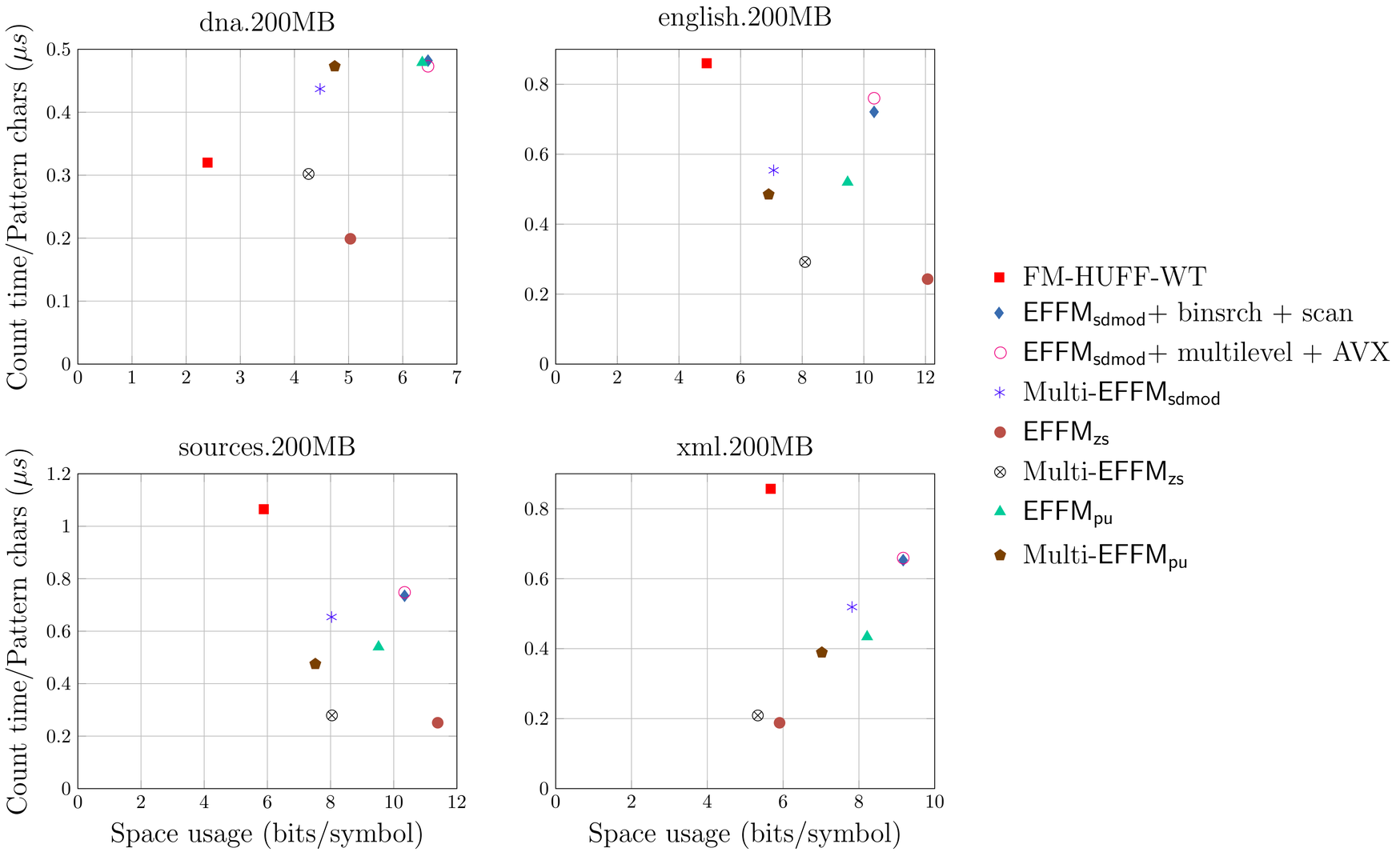


New Approaches: $\mathrm{EFFM}_{\mathrm{pu}}$

$B = 000001000000\ 100000010011\ 000010000000\ 010000100000\ 001100001100$ $\quad \Sigma = \{\$, i, m, p, s\}$
$l = 2$



New Approaches



Future Work

- Explore the use of our indexes for querying labelled trees and automata
- Exploit AVX instructions more
    - AVX still haven't been used much in succinct data structures
- Using our new EF implementations in other applications where EF is currently used

Thank you!