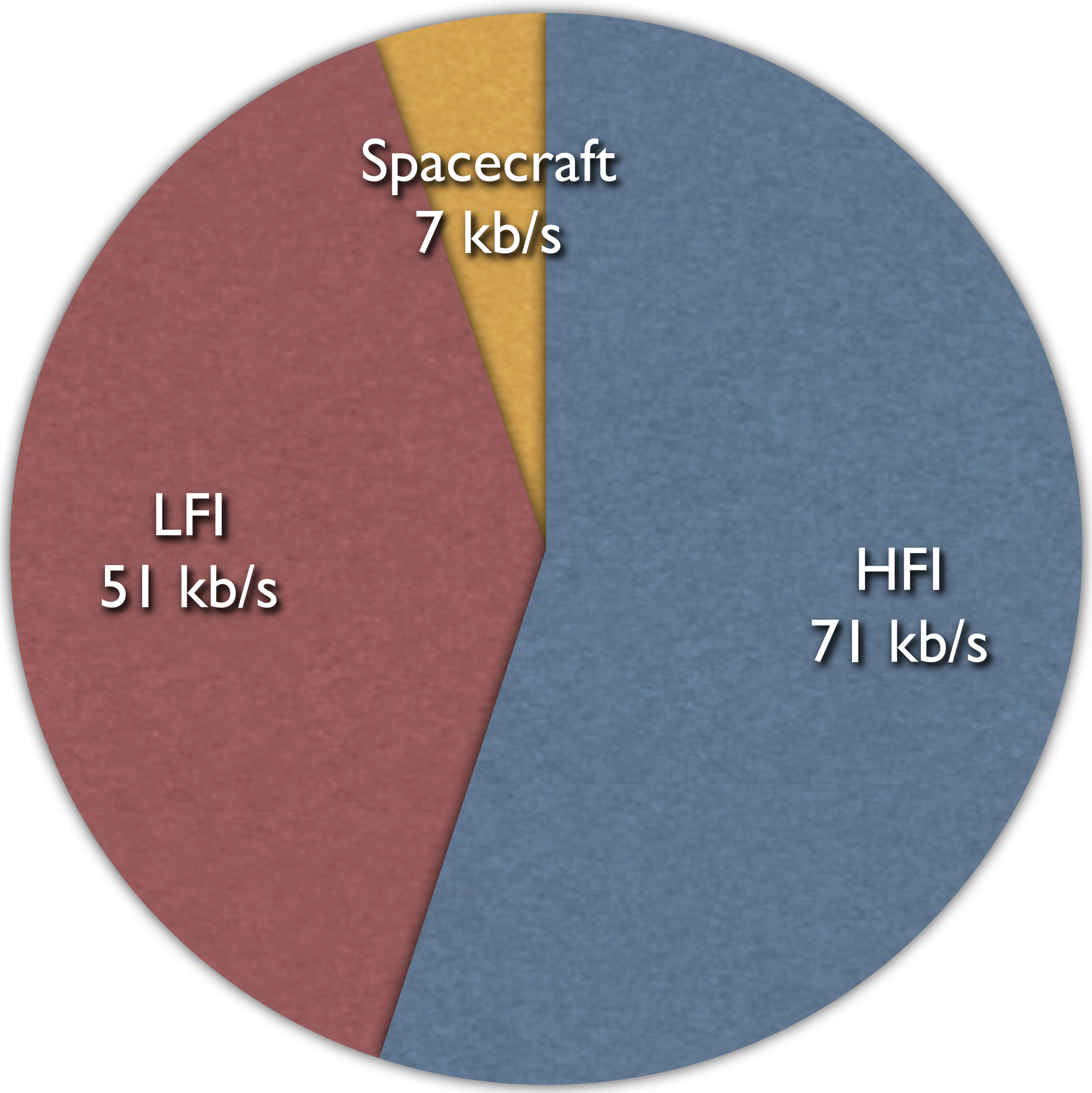
# Beyond PKZip: Data compression for Planck/LFI

Maurizio Tomasi

Astrosiesta INAF, April, 7th 2011

# Why do we need a compressor for LFI?




LFI data need ~2.4 times more than 51 kb/s!

# Part I
# Principles of data compression

# Types of compressors

The two categories of algorithms I am going to discuss today are:

- Dictionary compressors;
- Statistical compressors.

# Dictionary compressors: example

| When you find this… | …substitute it with this |
|---|---|
| as far as I know | AFAIK |
| by the way | BTW |
| id est | ie. |
| exempli gratia | eg. |
| post scriptum | ps. |

# Dictionary compressors

- Principle: substitute recurring patterns of symbols with shorter sequences
- Many compression programs belong to this family: PKZip, RAR, GZip, BZip2…
- This compression is very sensitive to the order of the symbols in the input!

# Statistical compressors

- Compressor of this type work on the bit sequences which encode each symbol.
- They compress better when some *symbols* (not patterns!) occur more than others.
- Popular example: the JPEG file format.

# An example

1304 1304 1301 1302 1301 1303 1304 1304 (16x8=128 bit)

# An example

1304 1304 1301 1302 1301 1303 1304 1304 (16x8=128 bit)

↓

11 11 00 01 00 10 11 11 (2x8=16 bit)

| Symbol | Bit mask |
|---|---|
| 1301 | 00 |
| 1302 | 01 |
| 1303 | 10 |
| 1304 | 11 |

# An example

11 11 00 01 00 10 11 11 (16 bit)

↓

0 0 10 110 10 111 0 0 (14 bit)

| Symbol | Frequency | Bit mask |
|---|---|---|
| 11 | 4 | 0 |
| 00 | 2 | 10 |
| 01 | 1 | 110 |
| 10 | 1 | 111 |

# Information entropy

- To estimate the performance of a statistical compressor, Shannon defined the so-called *information entropy*.
- If the *i*-th symbol occurs with frequency *p(i)*, then the information entropy of the signal is

$$H = -\sum_i p(i) \log_2 p(i)$$

(in our previous example, *H* = 7/4)

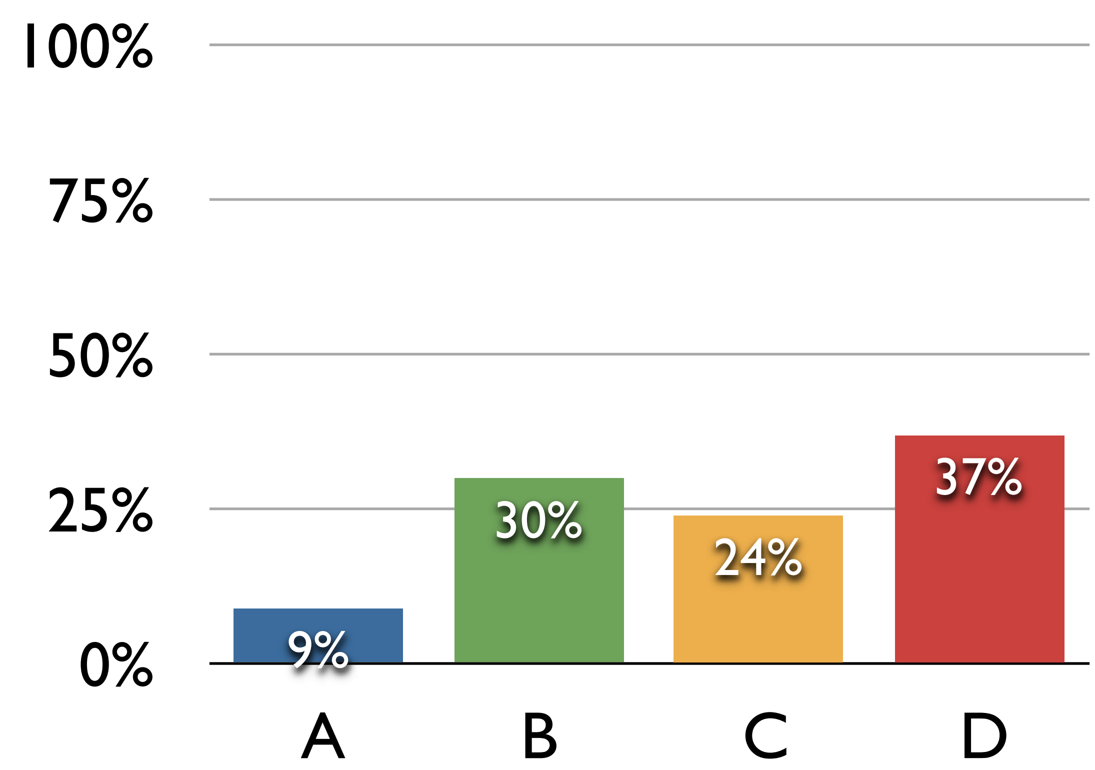
# Maximum achievable compression ratio

The maximum compression ratio achievable by any statistical compressor can be calculated analytically:

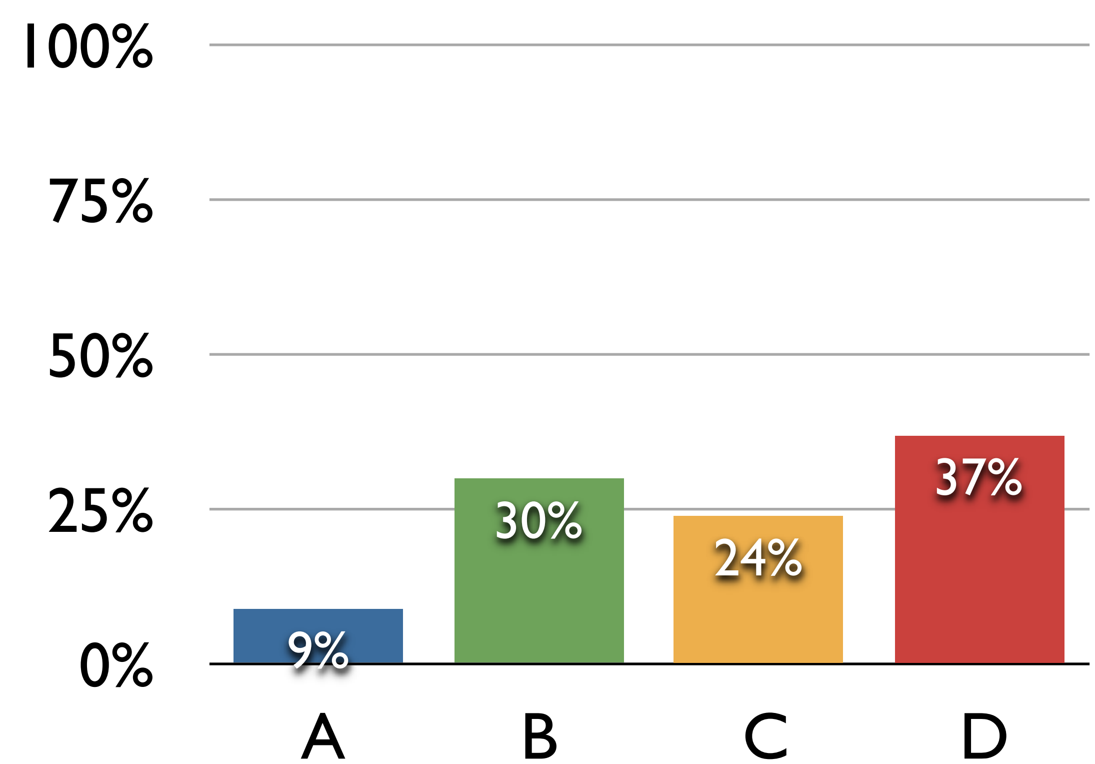
$$c_r^{\max} = \frac{n_{\mathrm{bits}}}{H} = \frac{\log_2 N_{\mathrm{symb}}}{H}$$

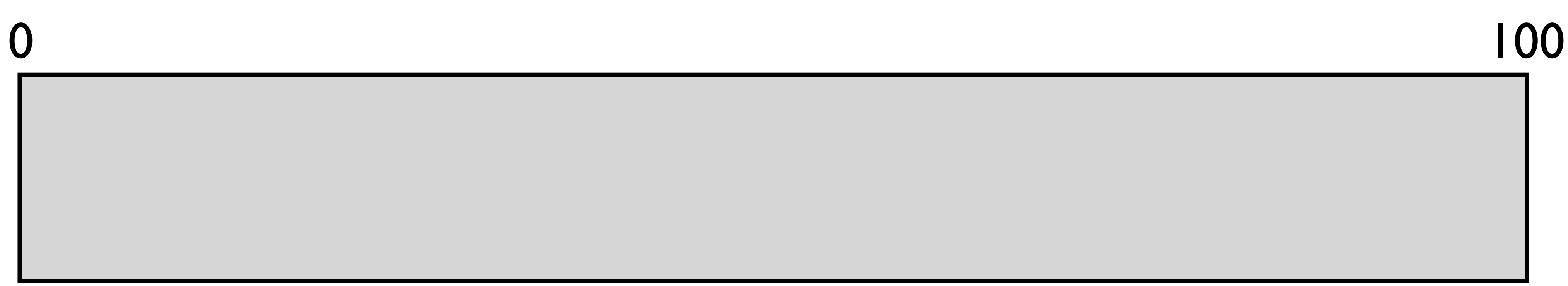
In our previous example we reached this limit (16/14 ~ 1.142). We can be proud!

# Arithmetic compression

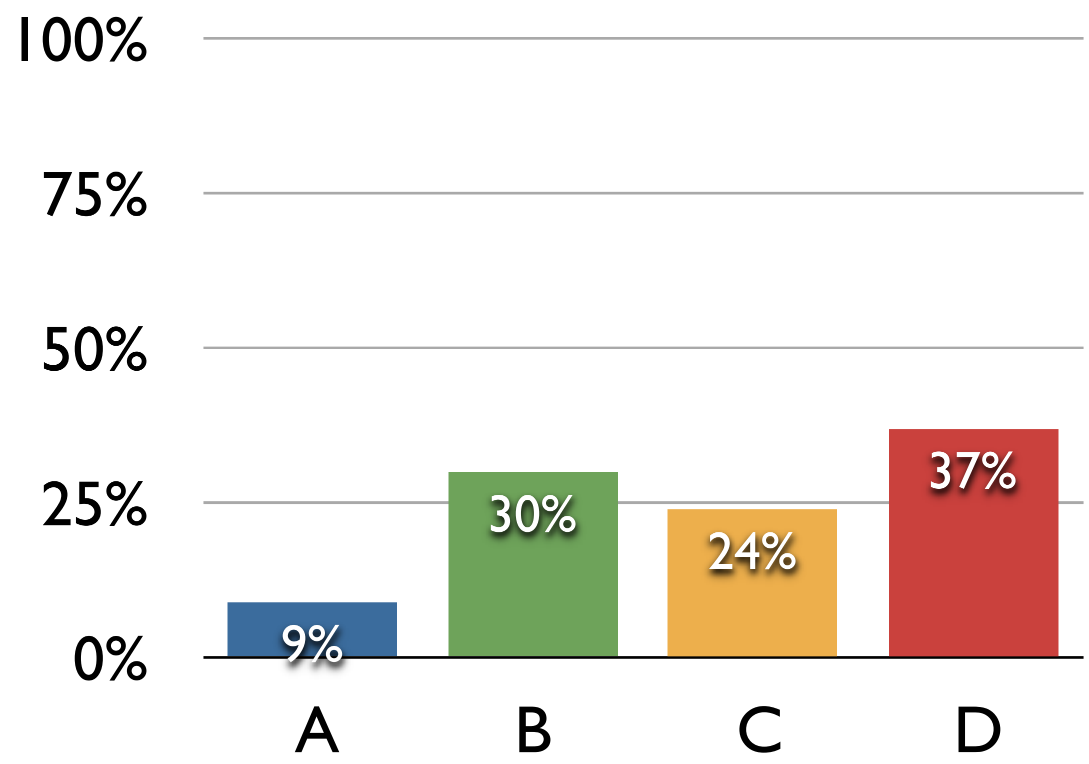


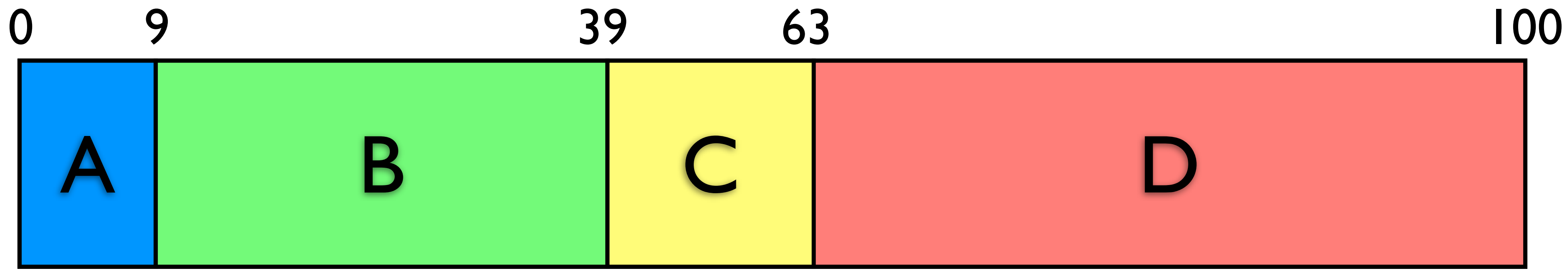
# Arithmetic compression



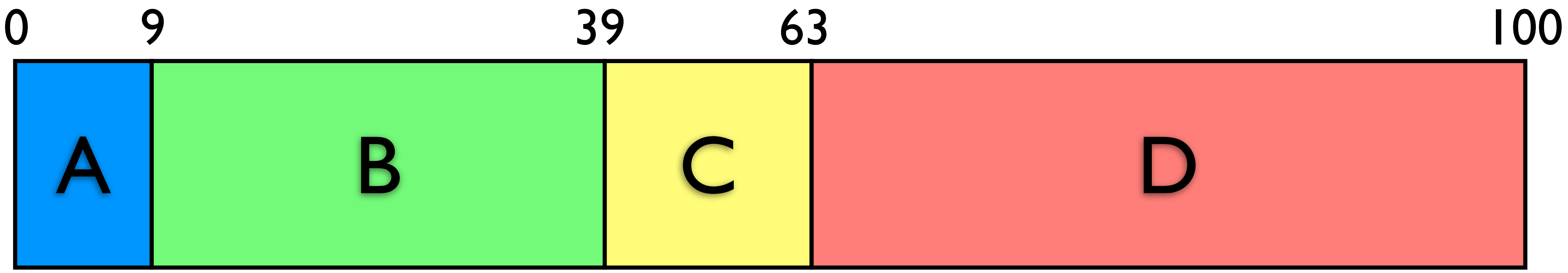



# Arithmetic compression

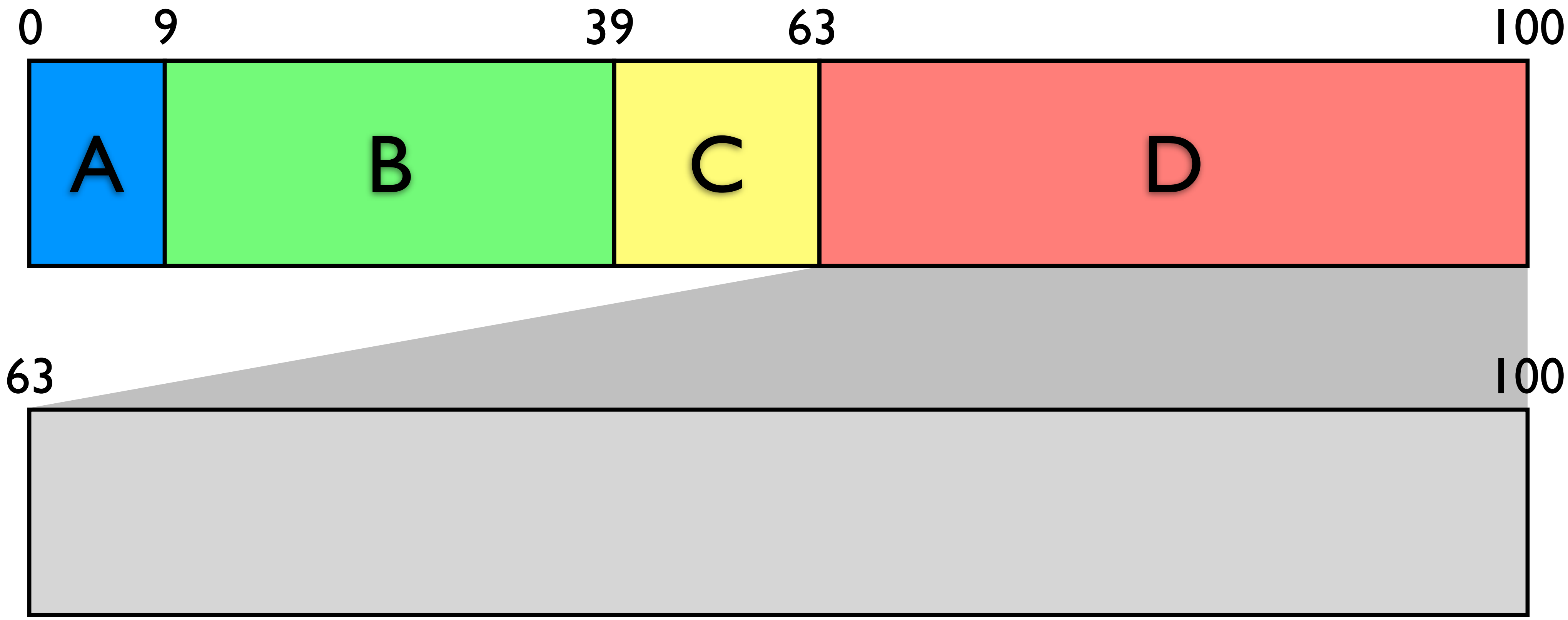


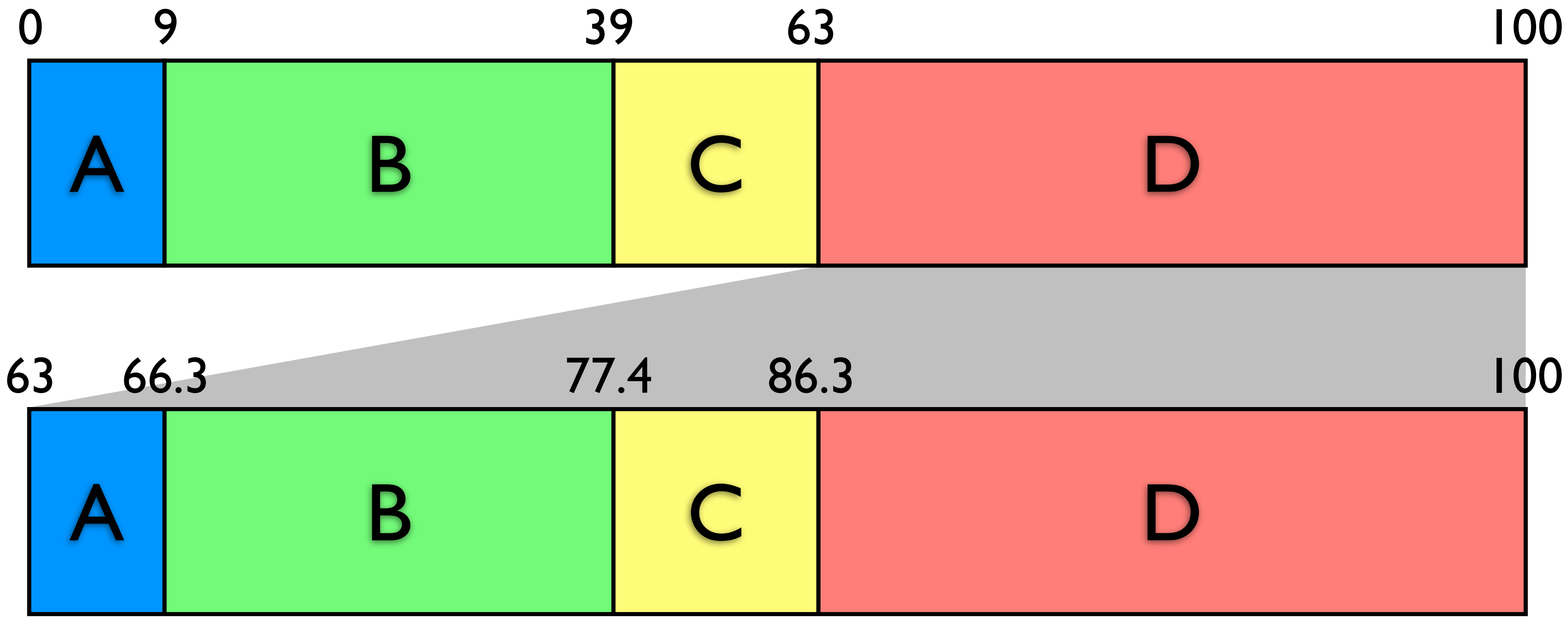
Input: D C B



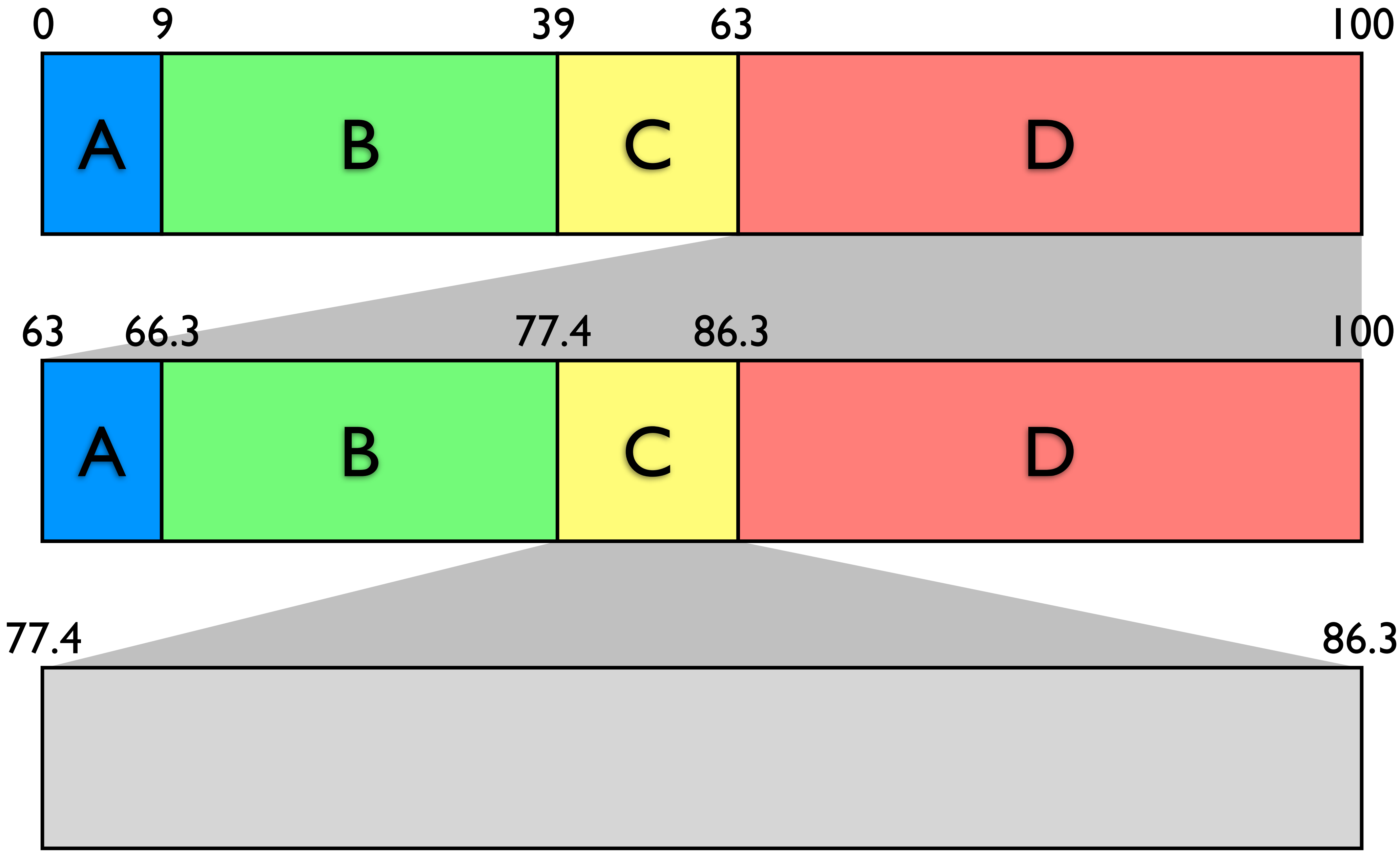
# Input: D C B



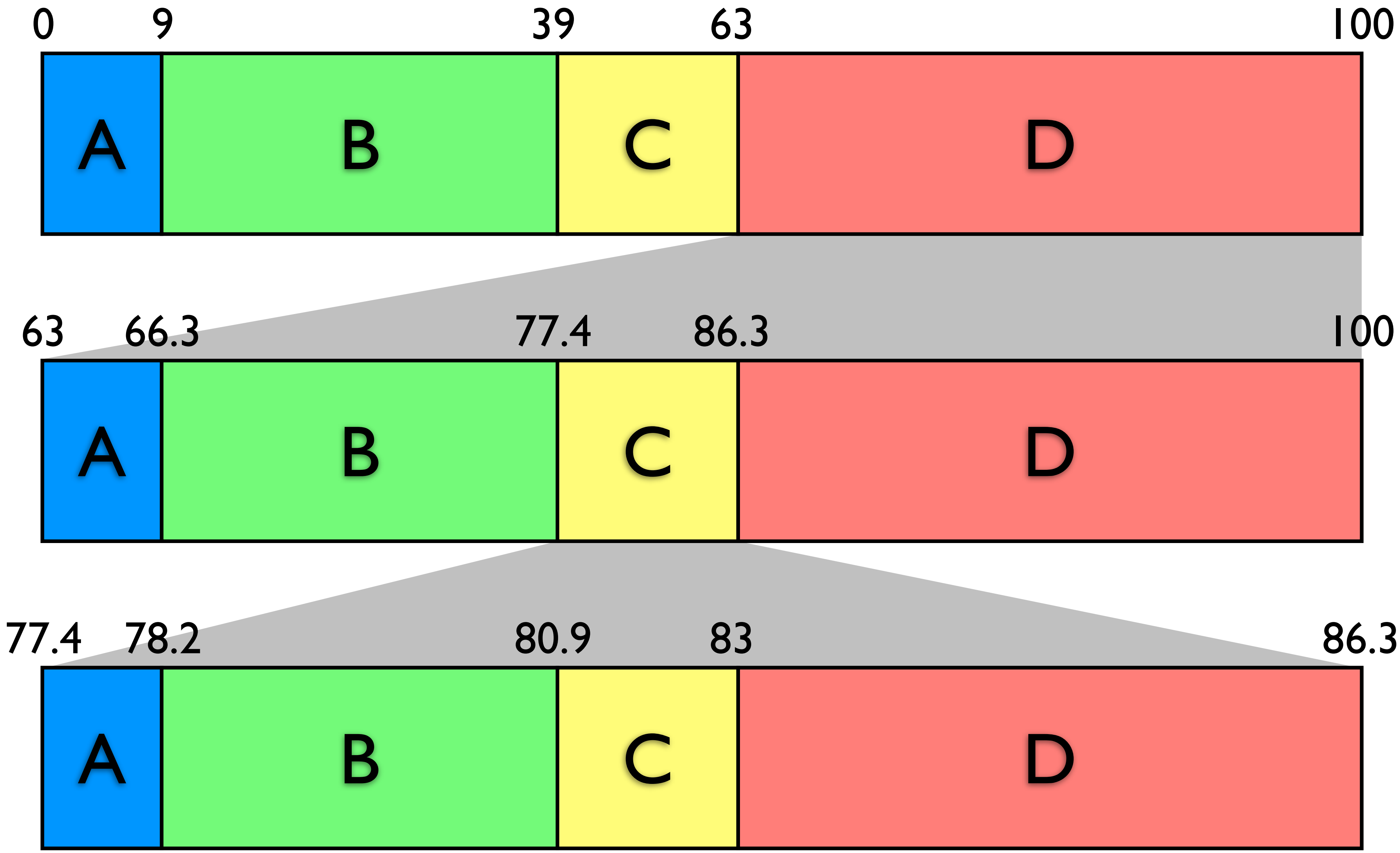
# Input: D C B



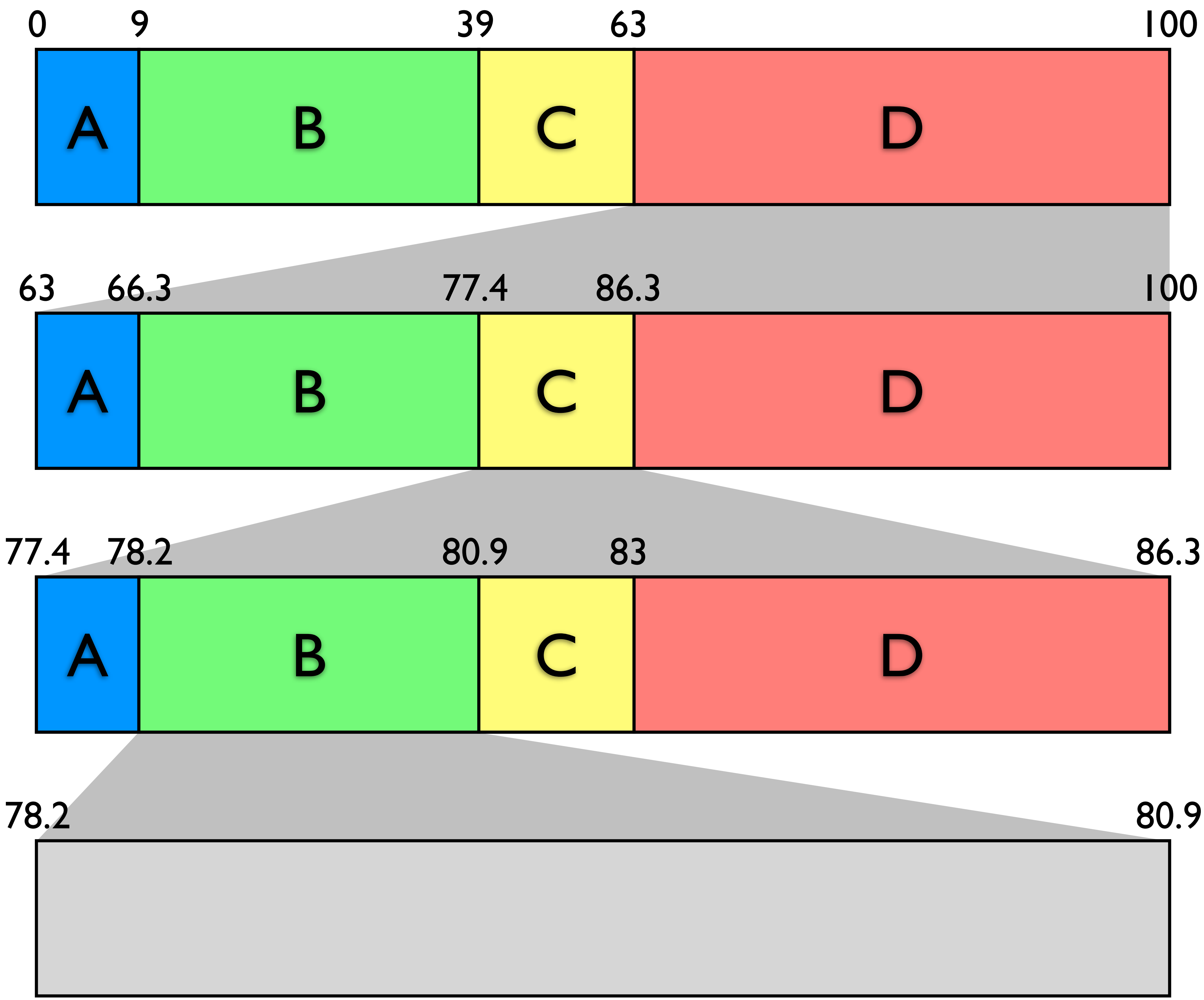
# Input: D C B



# Input: D C B

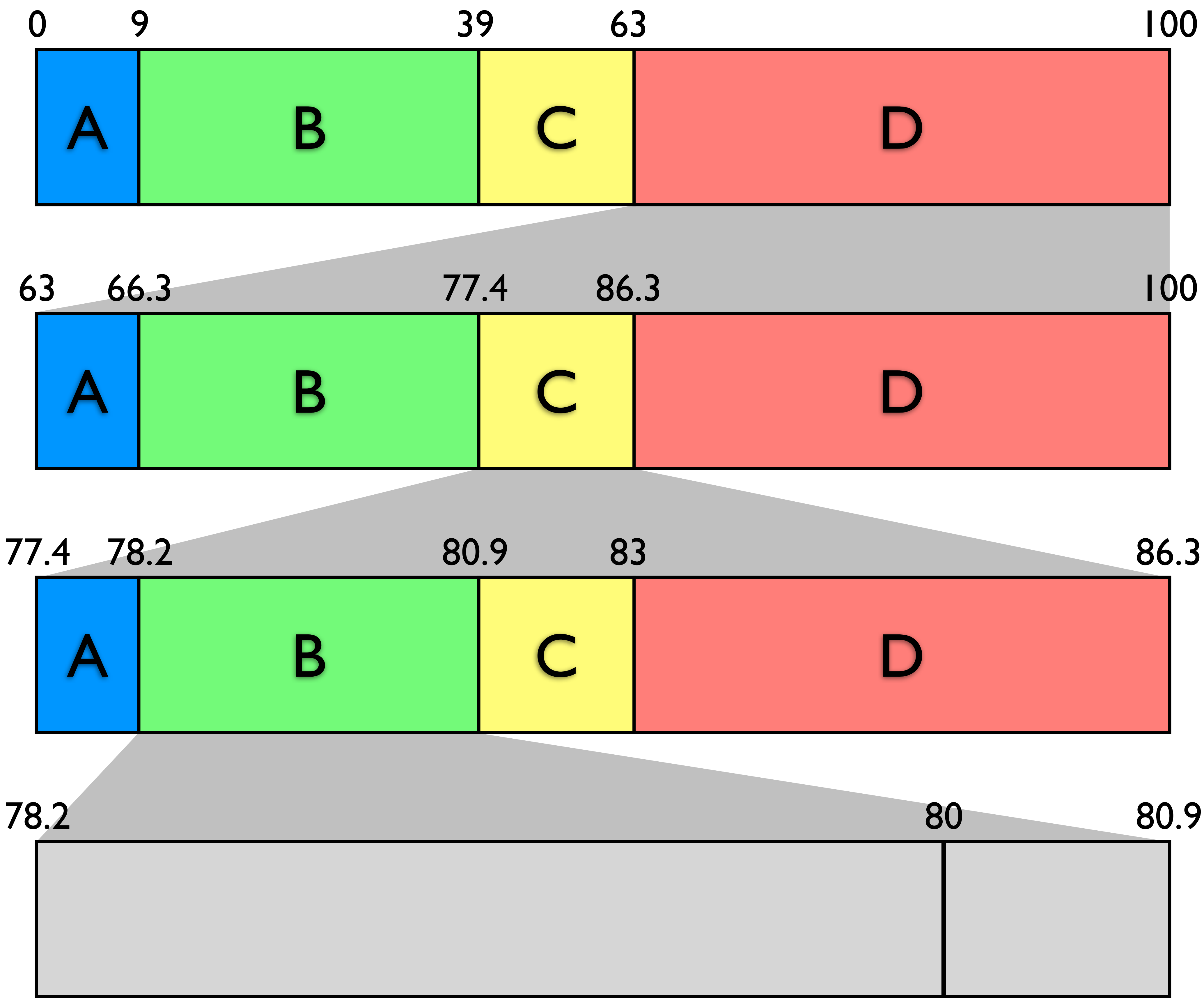


# Input: D C B



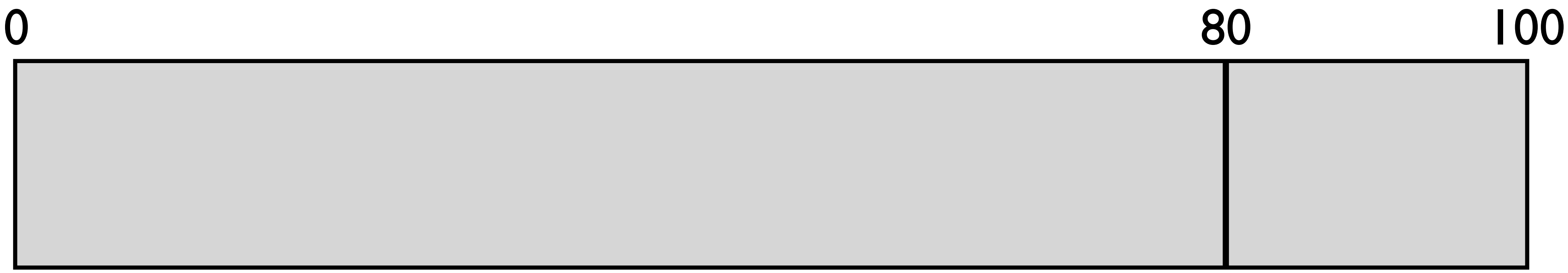
0
9
39
63
100
A
B
C
D
63
66.3
77.4
86.3
100
A
B
C
D
77.4
78.2
80.9
83
86.3
A
B
C
D
78.2
80.9

# Input: D C B

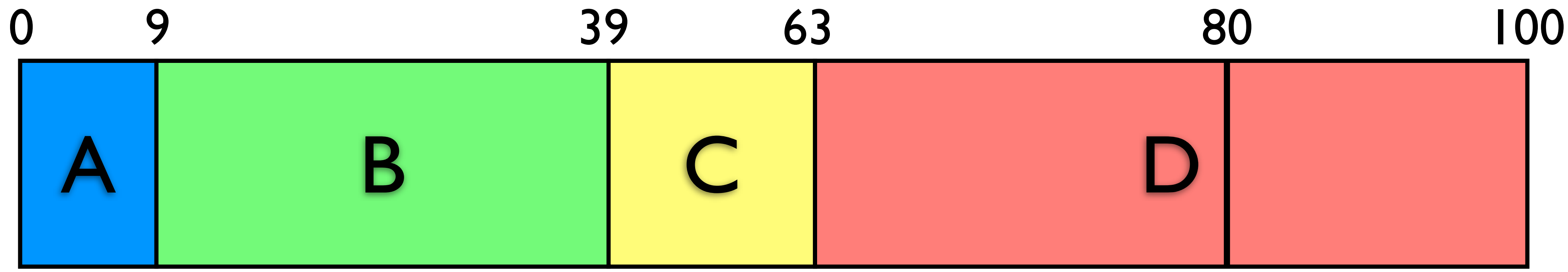


0
9
39
63
100
A
B
C
D
63
66.3
77.4
86.3
100
A
B
C
D
77.4
78.2
80.9
83
86.3
A
B
C
D
78.2
80
80.9

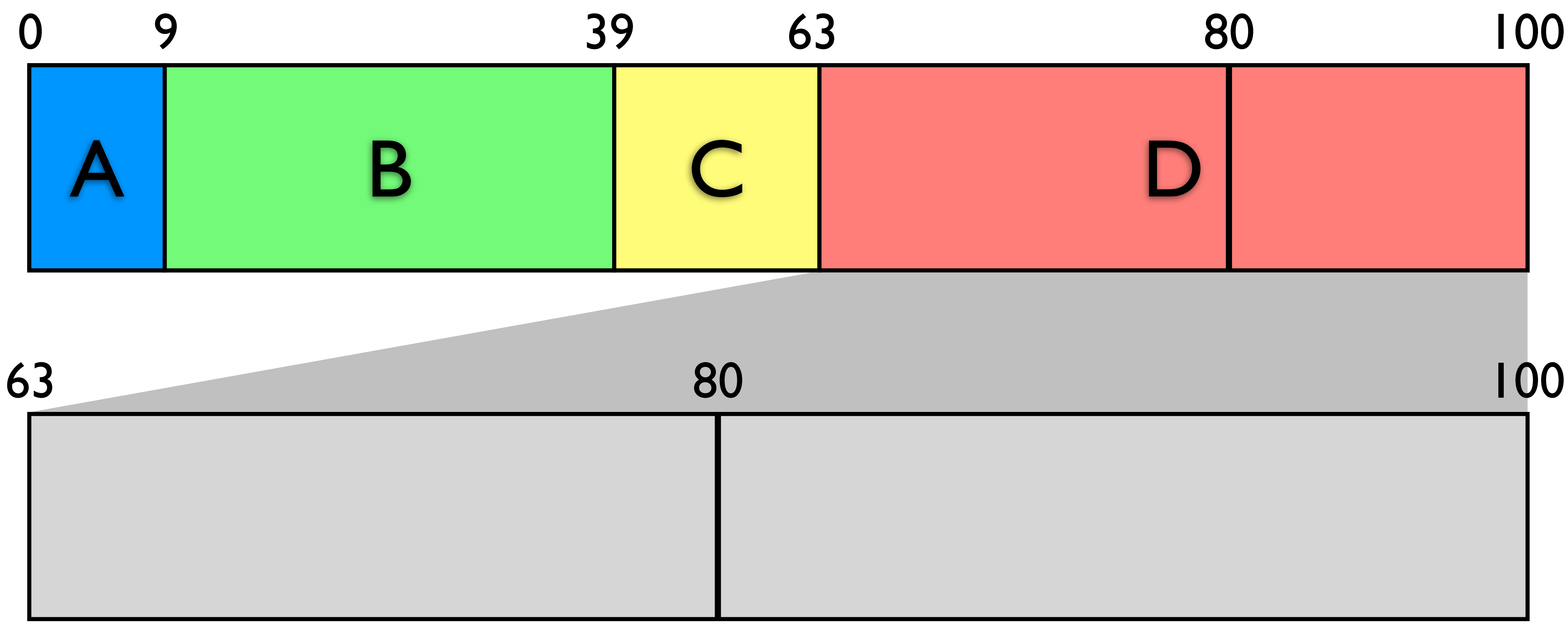
# Input: 80



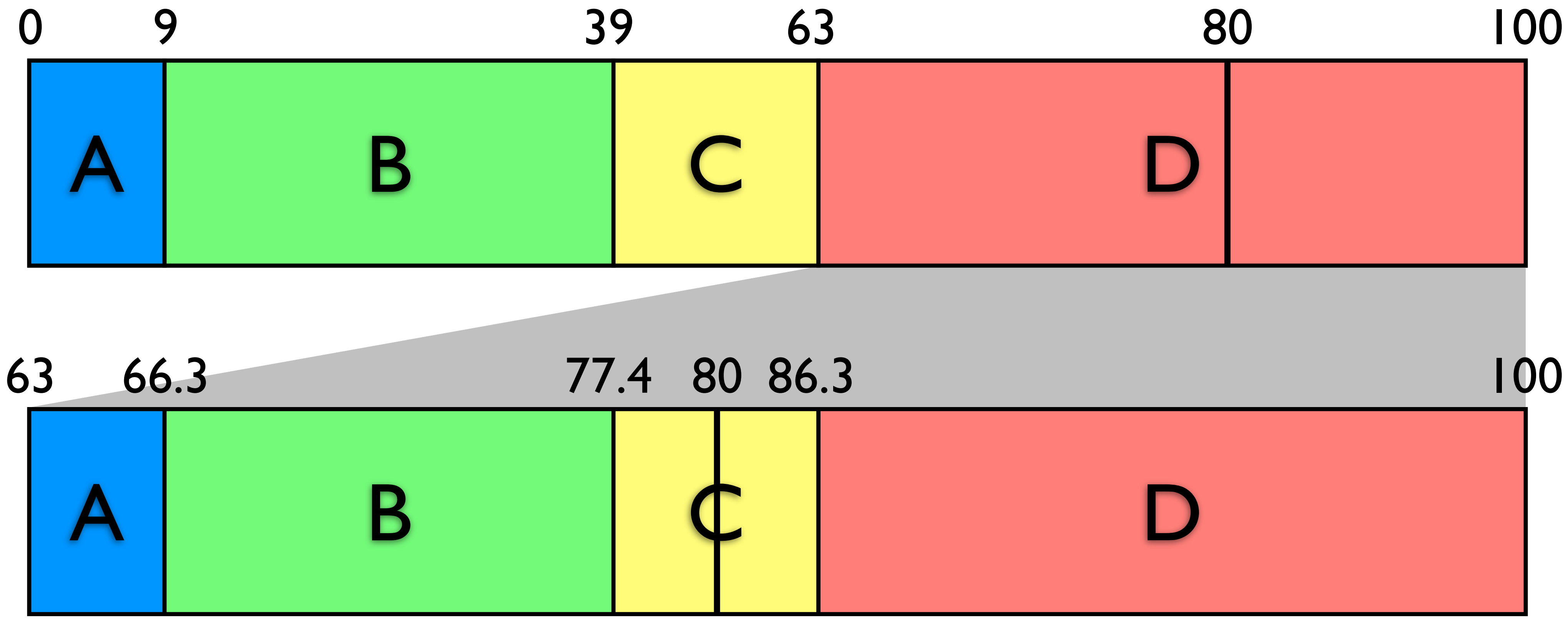
# Input: 80



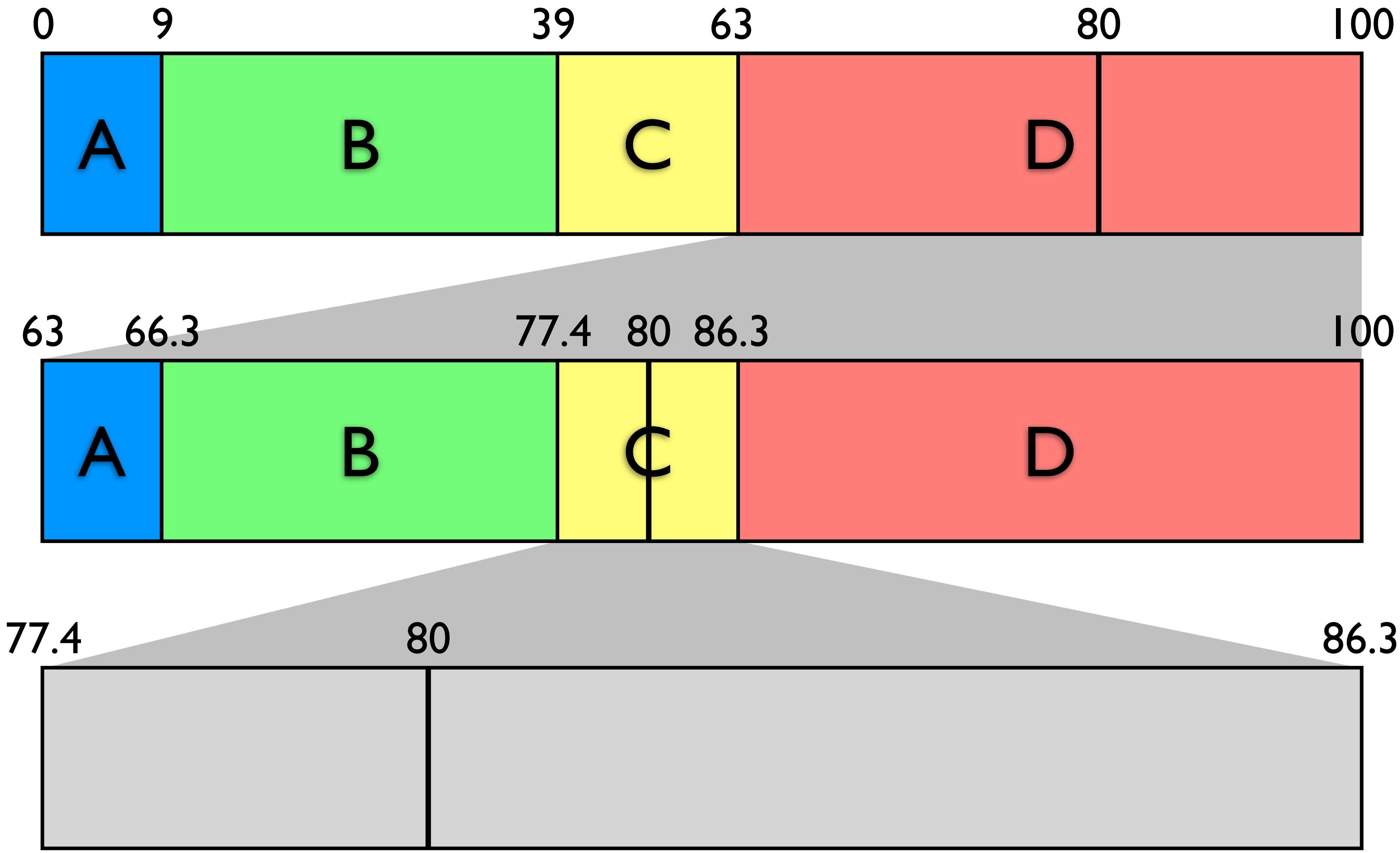
# Input: 80



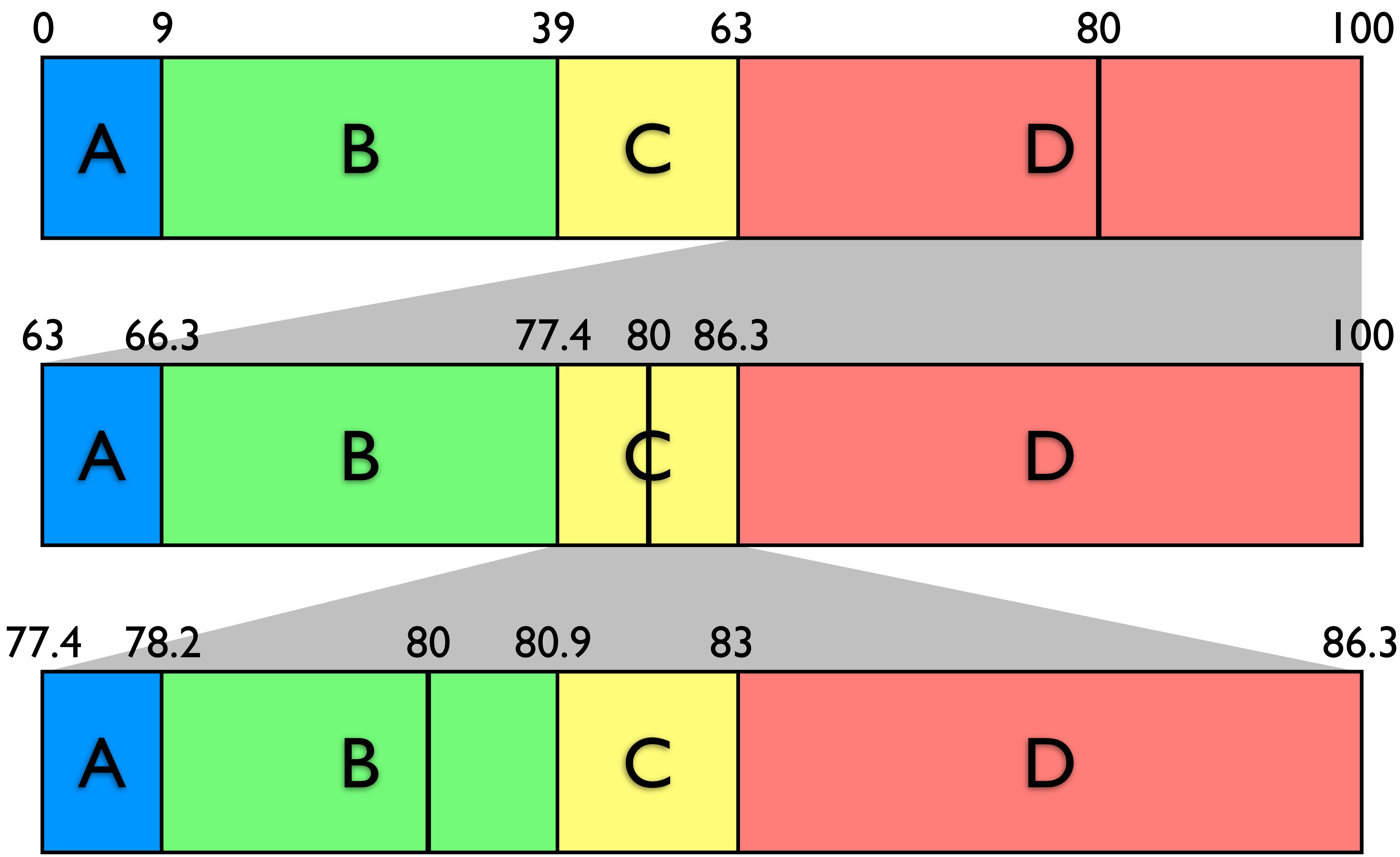
# Input: 80



# Input: 80

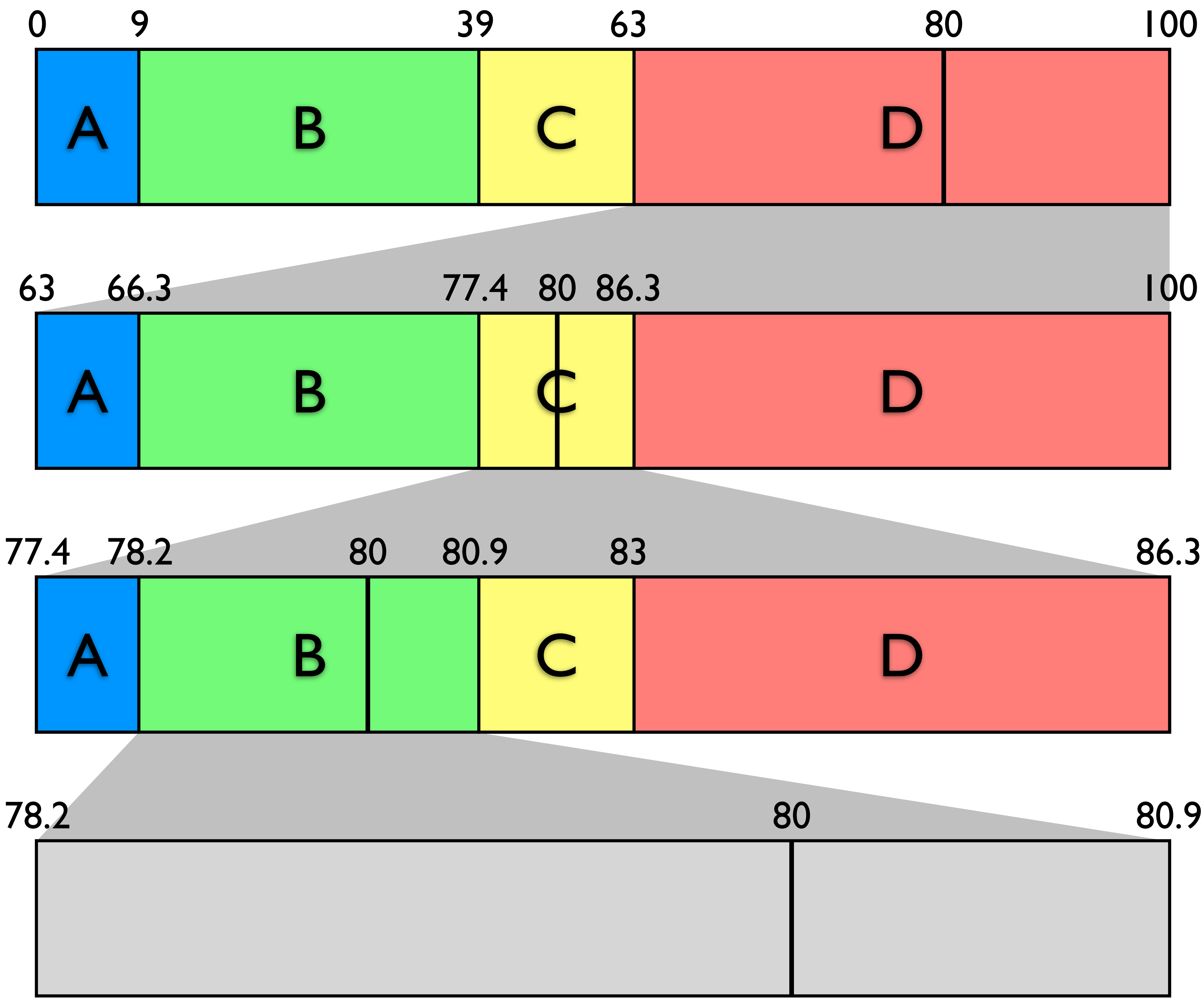


# Input: 80



0
9
39
63
80
100
A
B
C
D
63
66.3
77.4
80
86.3
100
A
B
C
D
77.4
78.2
80
80.9
83
86.3
A
B
C
D

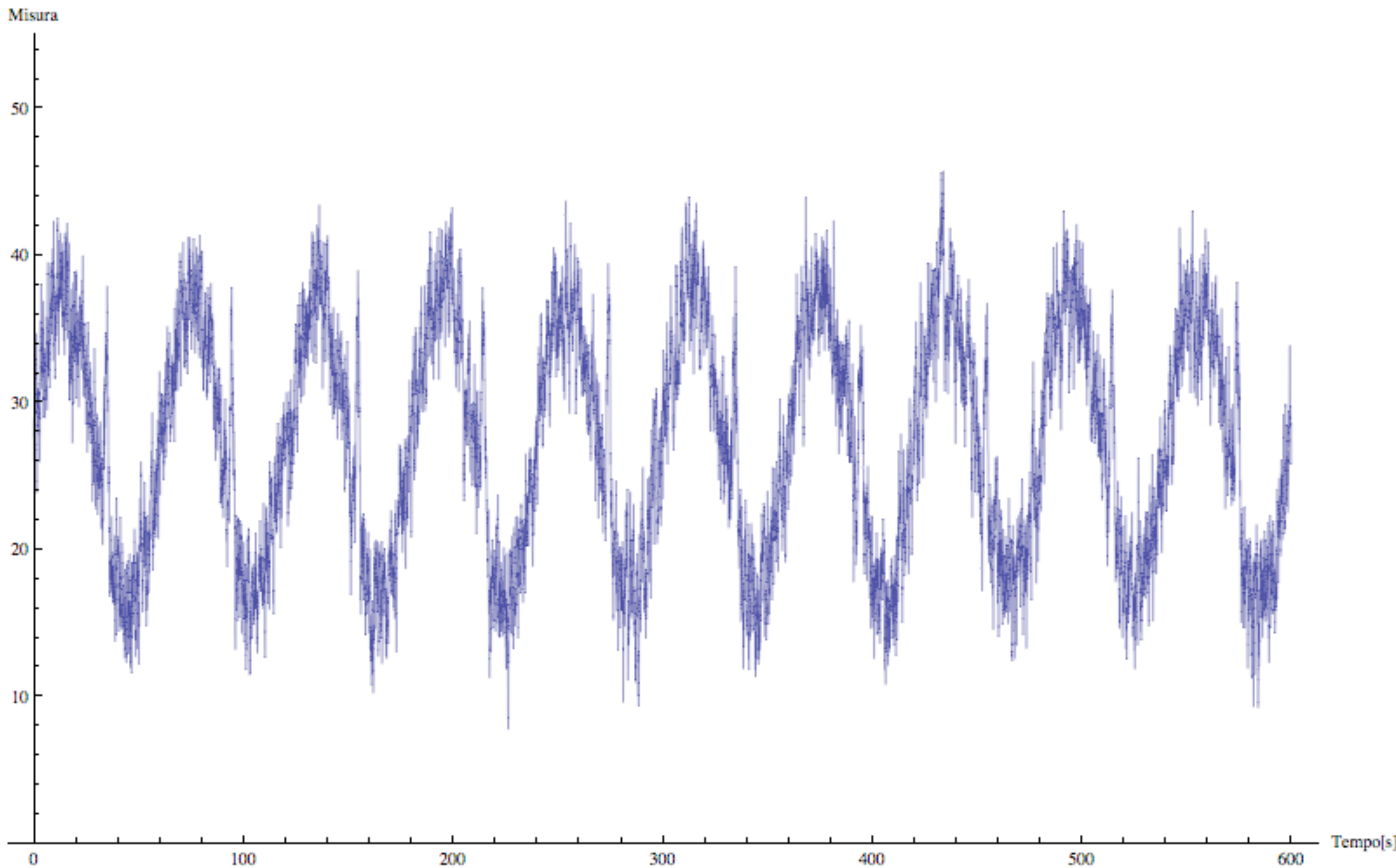
# Input: 80



0
9
39
63
80
100
A
B
C
D
63
66.3
77.4
80
86.3
100
A
B
C
D
77.4
78.2
80
80.9
83
86.3
A
B
C
D
78.2
80
80.9

# Part II
# The Planck/LFI case

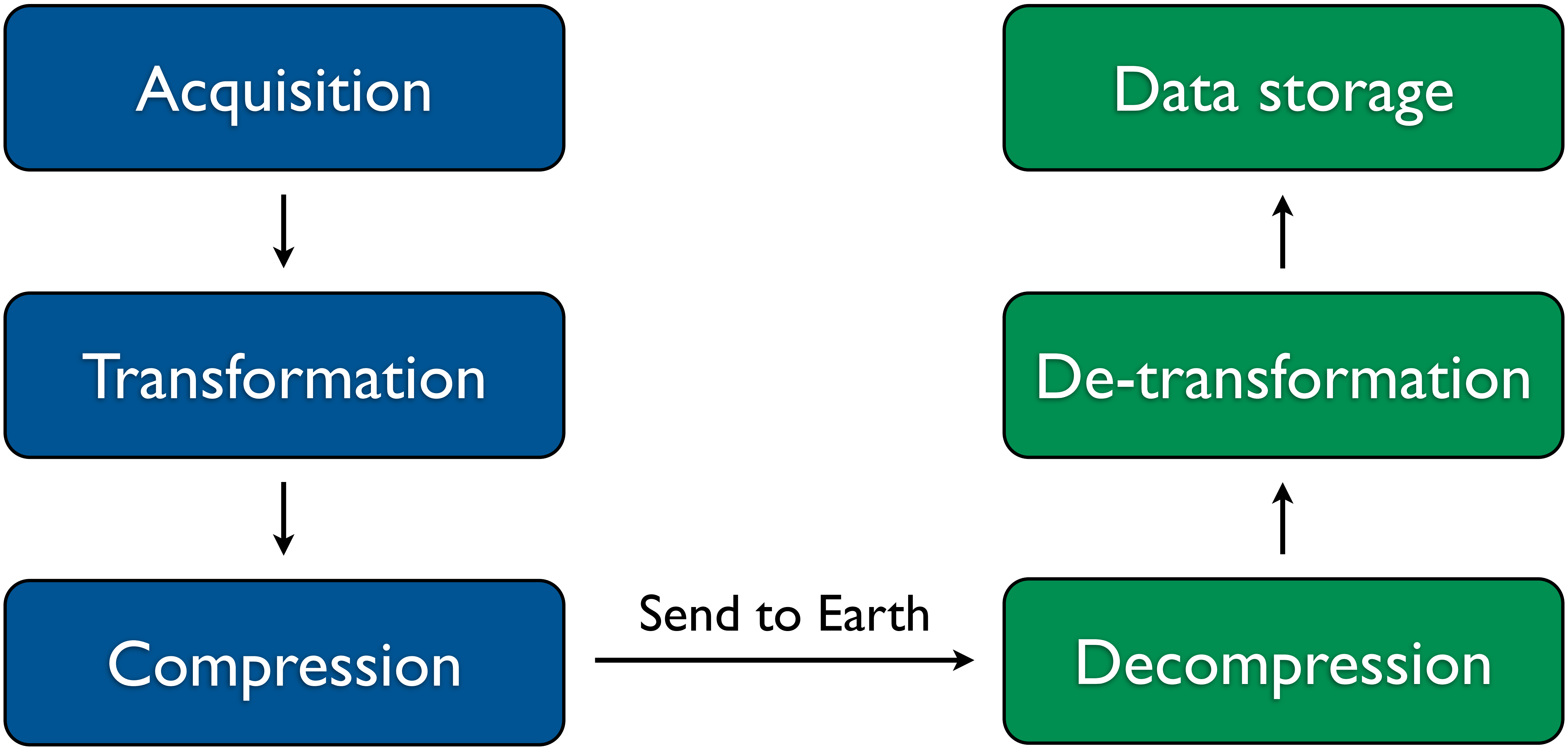
# Compressing LFI data



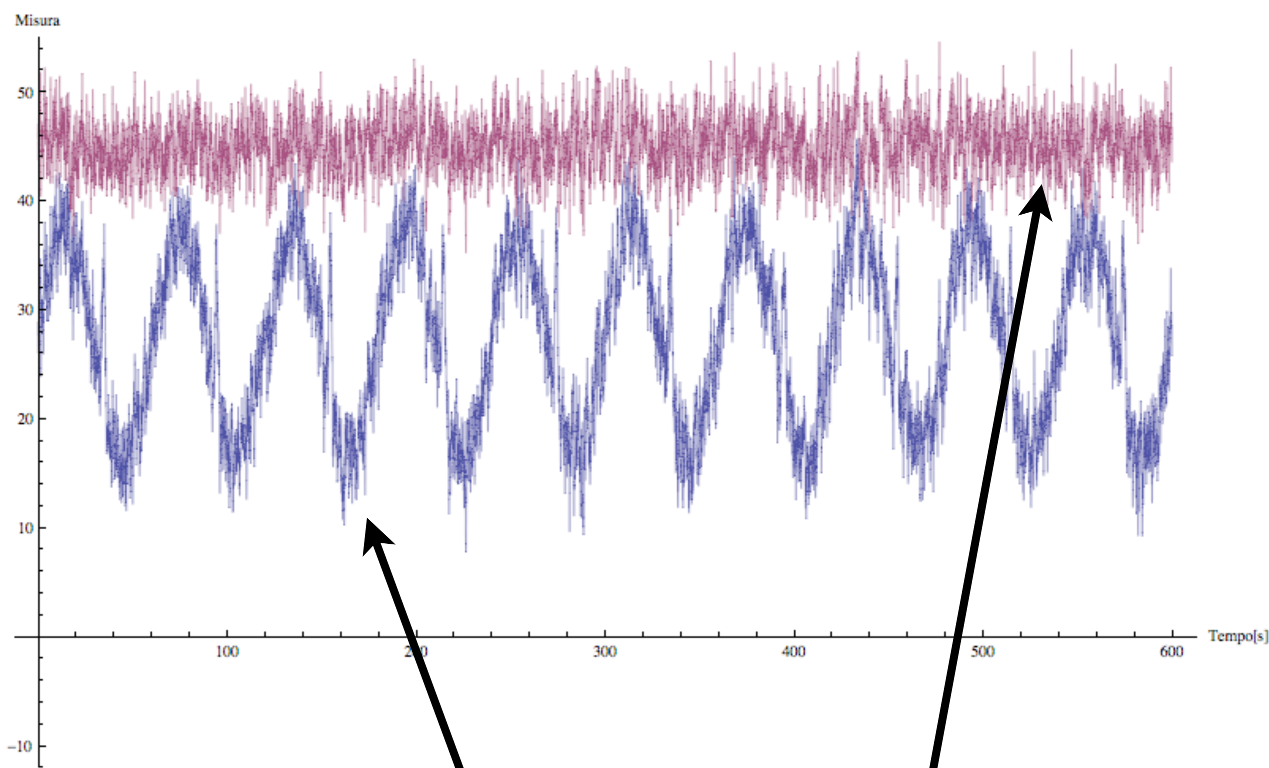
# What to use with radiometric output?

- Dictionary compressors are not useful here (no recurring patterns)
- Arithmetic compression (a statistical technique) is the best choice…
- …but it is not able to reach $c_r$~2.4: we need to decrease *H* before the compression!

# The LFI compression pipeline



# A non-reversible transform




$$q_1 = \mathrm{round}_{16}\Big( (x_s - r_1 x_r + \Delta) \times s \Big),$$

$$q_2 = \mathrm{round}_{16}\Big( (x_s - r_2 x_r + \Delta) \times s \Big)$$

# A non-reversible transform

- Differencing sky/ref reduces noise.
- The rounding operation increases the signal RMS, but in a way that *decreases H*!
- Requirement: such increment must be <10% of the intrinsic signal RMS.

# Optimisation and verification

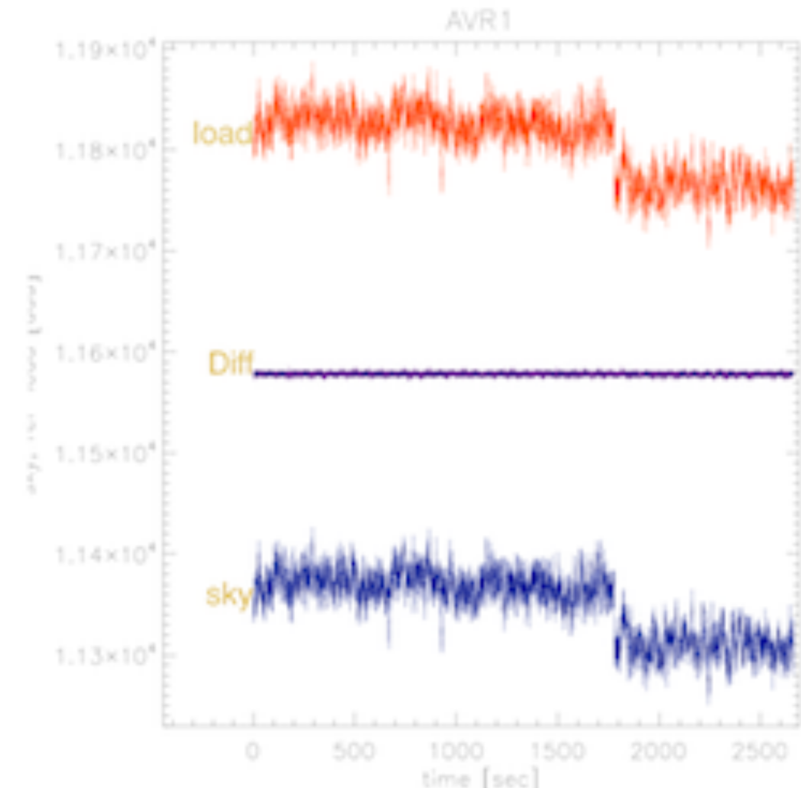
- To find the values of the parameters $r_1$, $r_2$, $\Delta$, $s$ for each detector (44 total), LFI sent *uncompressed* data to Earth for ~1 day.
- We analyzed such data on ground, found the best configuration and sent it back to LFI.
- During operations, we monitor the performance of the compressor daily.

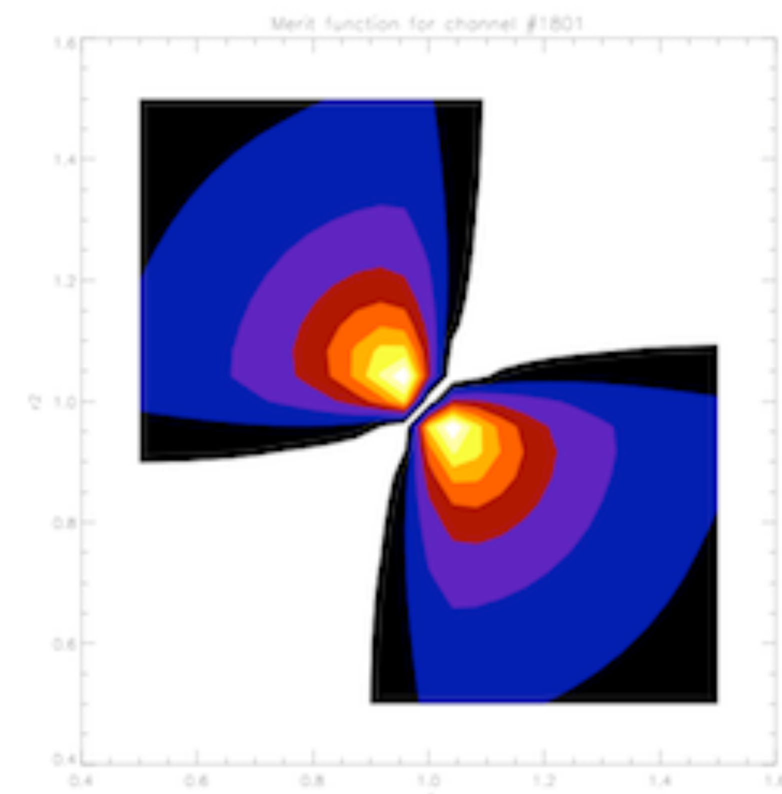
### Channel #1801

| | |
|---|---|
| **TOI statistics** | |
| Time Interval | [1626455700.0, 1626458400.0] s |
| Number of samples | 211874 |
| $V_{\text{sky}}$ mean ± rms | 11350.10 ± 32.44 ADU |
| $V_{\text{load}}$ mean ± rms | 11806.20 ± 33.56 ADU |
| $\sigma_{\text{sky,load}}$, $\varrho_{\text{sky,load}}$ | 11.9735 ADU$^2$, 0.0313 |
| $r$ | 0.9614 |
| **Optimized REBA Parameters** | |
| $N_{\text{aver}}$ | 52.0000 |
| $r_1$ | 1.0000 |
| $r_2$ | 0.9167 |
| $\mathcal{O}$ | -35.82 |
| $s_{\text{q}}$ | 2.9916 |
| Quack warning | no |
| **Processing** | |
| mean[$P_1$] + $\mathcal{O}\pm\sigma_1$ | $-491.99 \pm 2.34$ ADU |
| mean[$P_2$] + $\mathcal{O}\pm\sigma_2$ | $+491.86 \pm 2.58$ ADU |
| mean[$Q_1$]$\pm\sigma_{Q_1}$ | $-1471.83 \pm 7.02$ ADU |
| mean[$Q_2$]$\pm\sigma_{Q_1}$ | $+1471.43 \pm 7.74$ ADU |
| max($\lvert\mathcal{Q}_{\text{ack}}\rvert$) | 0.045929 |
| $C_{\text{r}}$ | $2.37 \pm 0.03$ (target: 2.40) |
| $\eta_{C_{\text{r}}}$ | $(82.49 \pm 0.49)$% |
| $H_{\text{pck}}$ | $5.57 \pm 0.05$ bits |
| $H_{\text{toi}}$ | 5.93 bits |
| $\epsilon_q$ | 1.5699 ADU (sky), 1.6366 ADU (ref), 0.0684 ADU (diff.) |
| $\frac{\epsilon_q}{\sigma}$ | 0.137859 (sky), 0.048750 (ref), 0.035396 (diff.) |
| **Telemetry** | |
| $N_{\text{pck}}$ | 414 |
| Uncomp. data length | 847496. bytes |
| Comp. data length | 356900 bytes |
| Comp. data rate | 1111.45 bits/sec |
| **Details of the analysis** | |
| Algorithm | `iterative` ('`diff`' strategy) |
| Grid $r_1 \times r_2$ | $[0.50, 1.50] \times [0.50, 1.50]$ (in $25 \times 25$ steps) |
| Elapsed Time | 60.5 s |

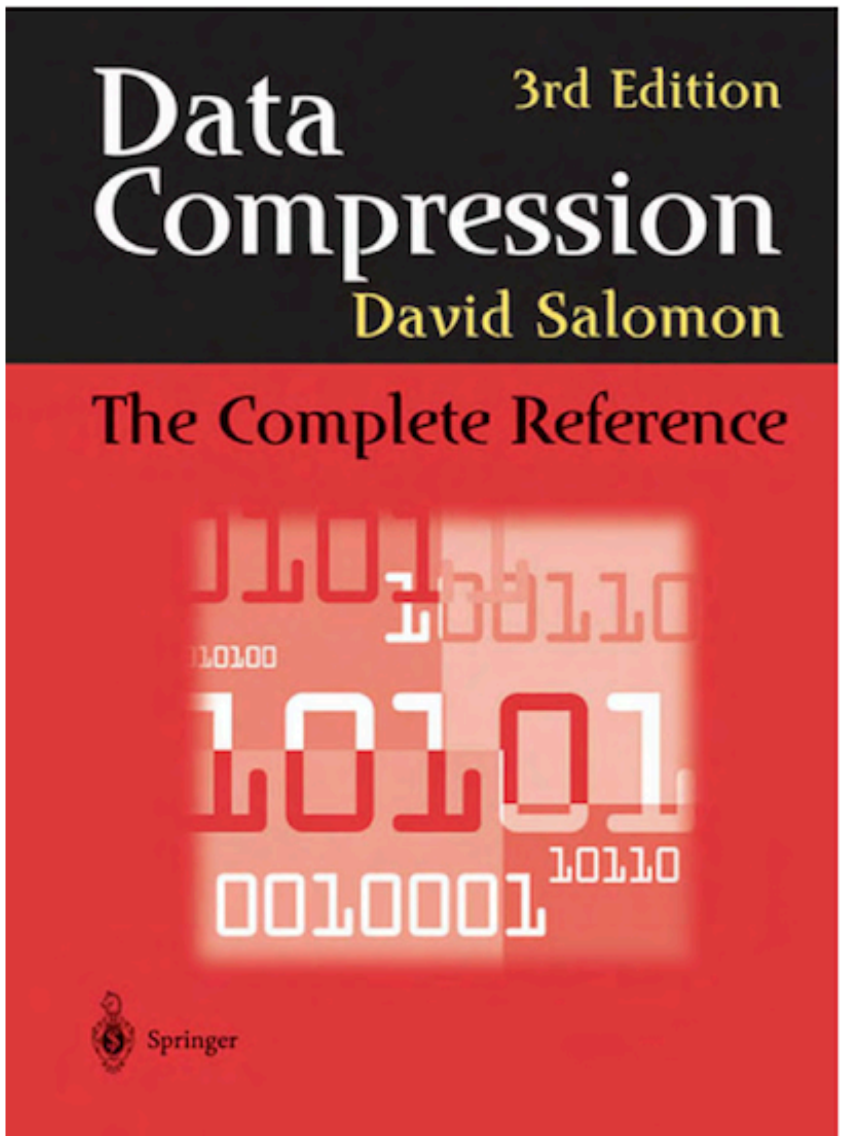


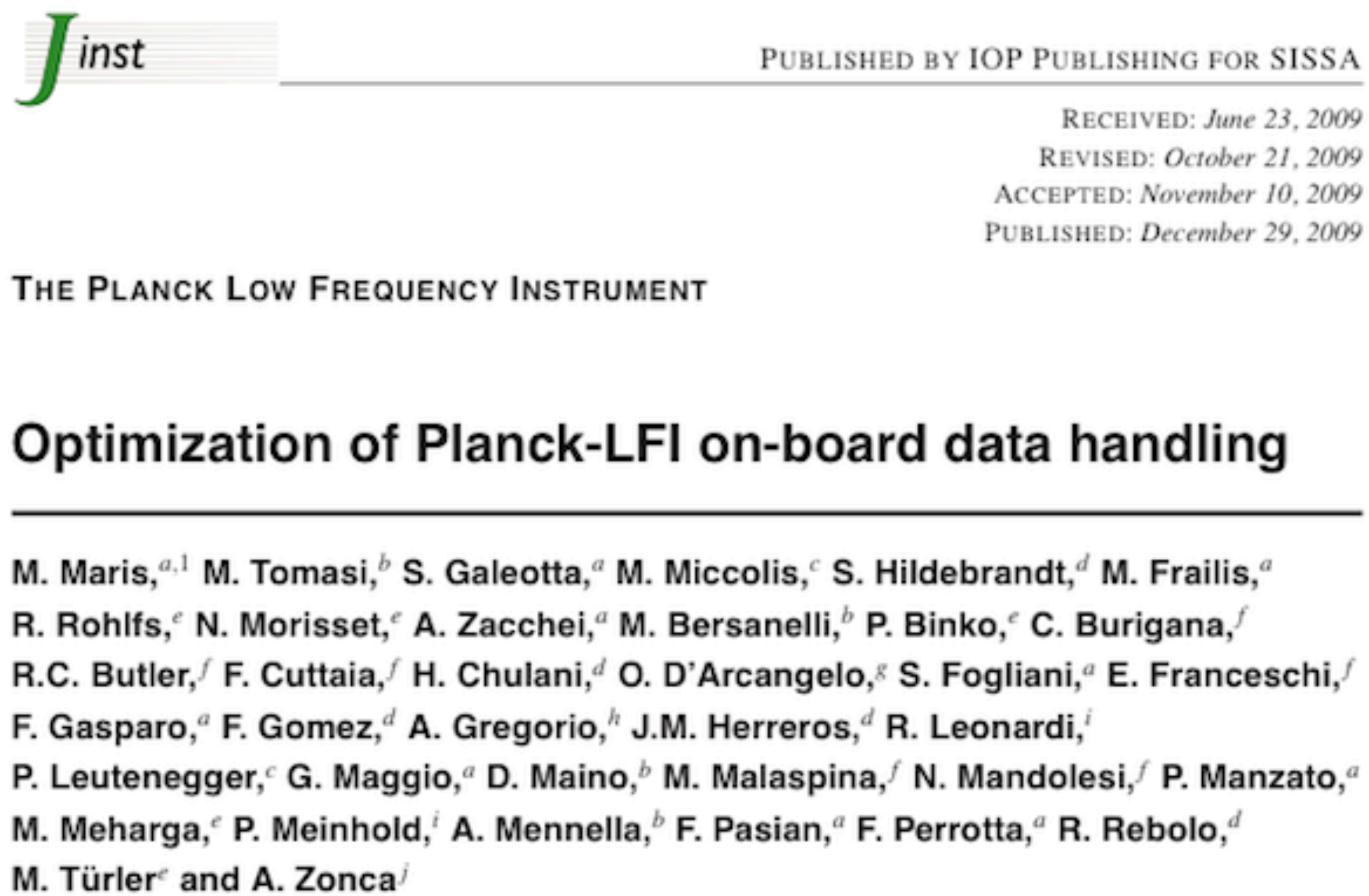




# Additional references



JINST

PUBLISHED BY IOP PUBLISHING FOR SISSA

RECEIVED: *June 23, 2009*
REVISED: *October 21, 2009*
ACCEPTED: *November 10, 2009*
PUBLISHED: *December 29, 2009*

THE PLANCK LOW FREQUENCY INSTRUMENT

**Optimization of Planck-LFI on-board data handling**

M. Maris,[a,1] M. Tomasi,[b] S. Galeotta,[a] M. Miccolis,[c] S. Hildebrandt,[d] M. Frailis,[a] R. Rohlfs,[e] N. Morisset,[e] A. Zacchei,[a] M. Bersanelli,[b] P. Binko,[e] C. Burigana,[f] R.C. Butler,[f] F. Cuttaia,[f] H. Chulani,[d] O. D'Arcangelo,[g] S. Fogliani,[a] E. Franceschi,[f] F. Gasparo,[a] F. Gomez,[d] A. Gregorio,[h] J.M. Herreros,[d] R. Leonardi,[i] P. Leutenegger,[c] G. Maggio,[a] D. Maino,[b] M. Malaspina,[f] N. Mandolesi,[f] P. Manzato,[a] M. Meharga,[e] P. Meinhold,[i] A. Mennella,[b] F. Pasian,[a] F. Perrotta,[a] R. Rebolo,[d] M. Türler[e] and A. Zonca[j]