情報の単位

情報の単位

- 情報の単位＝データの量（サイズ）の大きさを表す単位。データの量≠情報量（エントロピー、別の概念）
- ビット（bit）：2進数の1桁で表せる情報の単位。コンピュータで扱うことのできる情報の最小の単位。1ビットは0, 1に対応する二つの状態（on/off, あり／なし，白／黒，…）を表すことができる
- バイト（byte）：8ビットを表す単位。1バイト＝8ビット。前の符号付き2進数の例は1バイトの場合である
- 語／ワード（word）：コンピュータが扱うデータの単位。1語の大きさ（語長）は機種によって異なる。たとえばMIPSというCPUであれば、データも命令も1語を32ビット（4バイト）で構成している

情報の単位

	SI接頭辞	
	記号	換算
キロバイト(ビット)	kB(kb)	1kB(kb)=10^3byte(bit)=1000byte(bit)
メガバイト（ビット）	MB(Mb)	1MB(Mb)=1000kB(kb)
ギガバイト（ビット）	GB(Gb)	1GB(Gb)=1000MB(Mb)
テラバイト（ビット）	TB(Tb)	1TB(Tb)=1000GB(Gb)
ペタバイト（ビット）	PB(Pb)	1PB(Pb)=1000TB(Tb)

SI接頭辞
SI:（仏）Système International d'unités、（英）International System of Units。フランス語由来。国際単位系。時間，長さ，質量等と統一されている

情報の単位

	2進接頭辞	
	記号	換算
キビバイト(ビット)	KiB(Kib)	1KiB(Kib)=2^{10}byte(bit)=1024byte(bit)
メビバイト(ビット)	MiB(Mib)	1MiB(Mib)=2^{10}KiB(Kib)=1024KiB(Kib)
ギビバイト(ビット)	GiB(Gib)	1GiB(Gib)=2^{10}MiB(Mib)=1024MiB(Mib)
テビバイト(ビット)	TiB(Tib)	1TiB(Tib)=2^{10}GiB(Gib)=1024GiB(Gib)
ペビバイト(ビット)	PiB(Pib)	1PiB(Pib)=2^{10}TiB(Tib)=1024TiB(Tib)

2進接頭辞
IEC(International Electrotechnical Commission, 国際電気標準会議)によって定められたもの

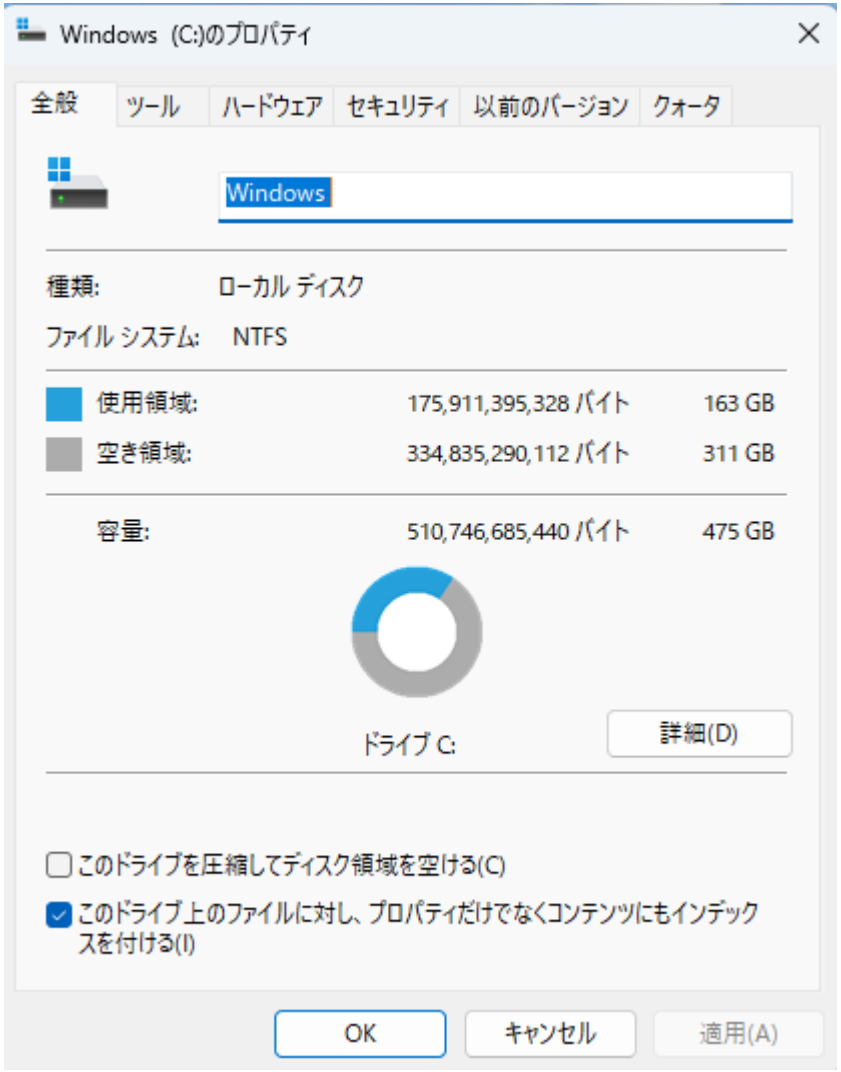
SI接頭辞 vs. 2進接頭辞

- （たとえばubuntuでdf –hで容量を調べると）計算機の容量表示はこの1024を基本とした2進接頭辞を使っている
- 一方、メーカのホームページや仕様書などにはSI接頭辞で記載する
- したがって、手元の計算機で容量を調べると、値は仕様書に比べ減っているように見える

例：Windows 11 PCのディスク容量の表示

- 購入時の仕様書では：512GB SSD

- しかしWindows (C:)のプロパティでは右図の通り

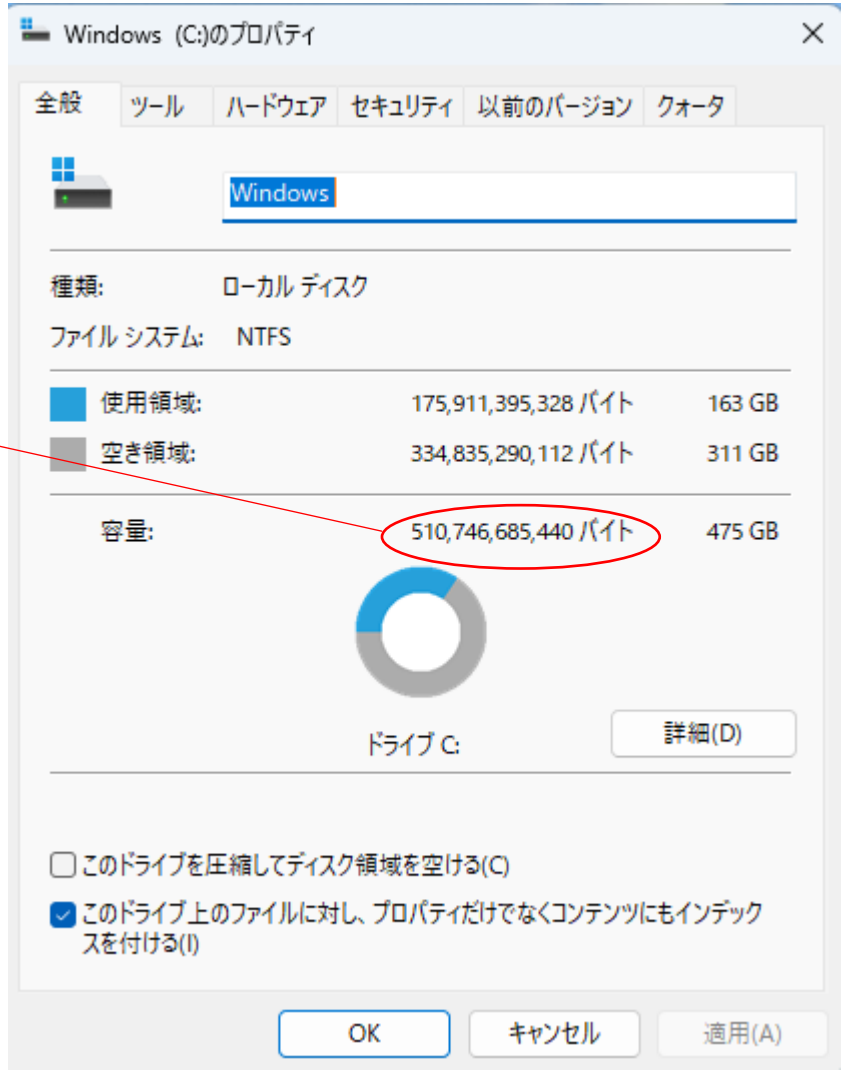


例：Windows 11 PCのディスク容量の表示

- 購入時の仕様書では：512GB SSD

SI接頭辞使用

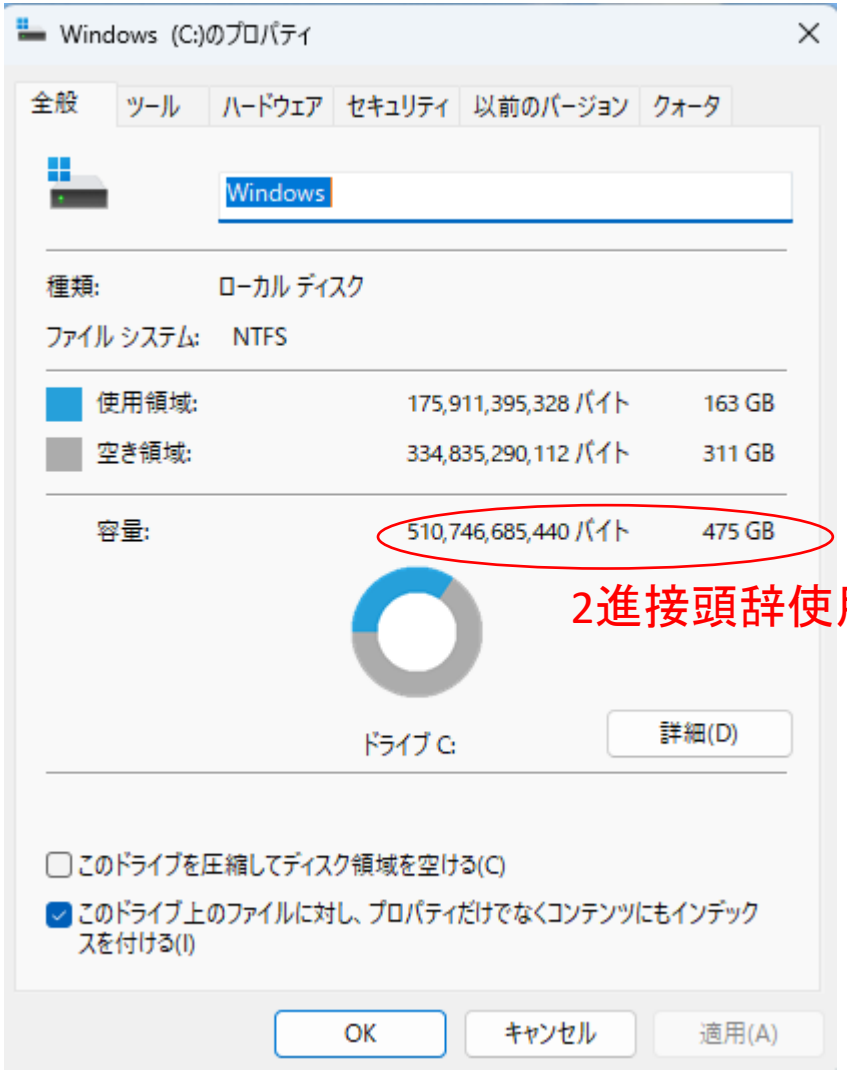
- しかしWindows (C:)のプロパティでは右図の通り



例：Windows 11 PCのディスク容量の表示

- 購入時の仕様書では：512GB SSD

- しかしWindows (C:)のプロパティでは右図の通り



例：Windows 11 PCのディスク容量の表示

- 購入時の仕様書では：512GB SSD

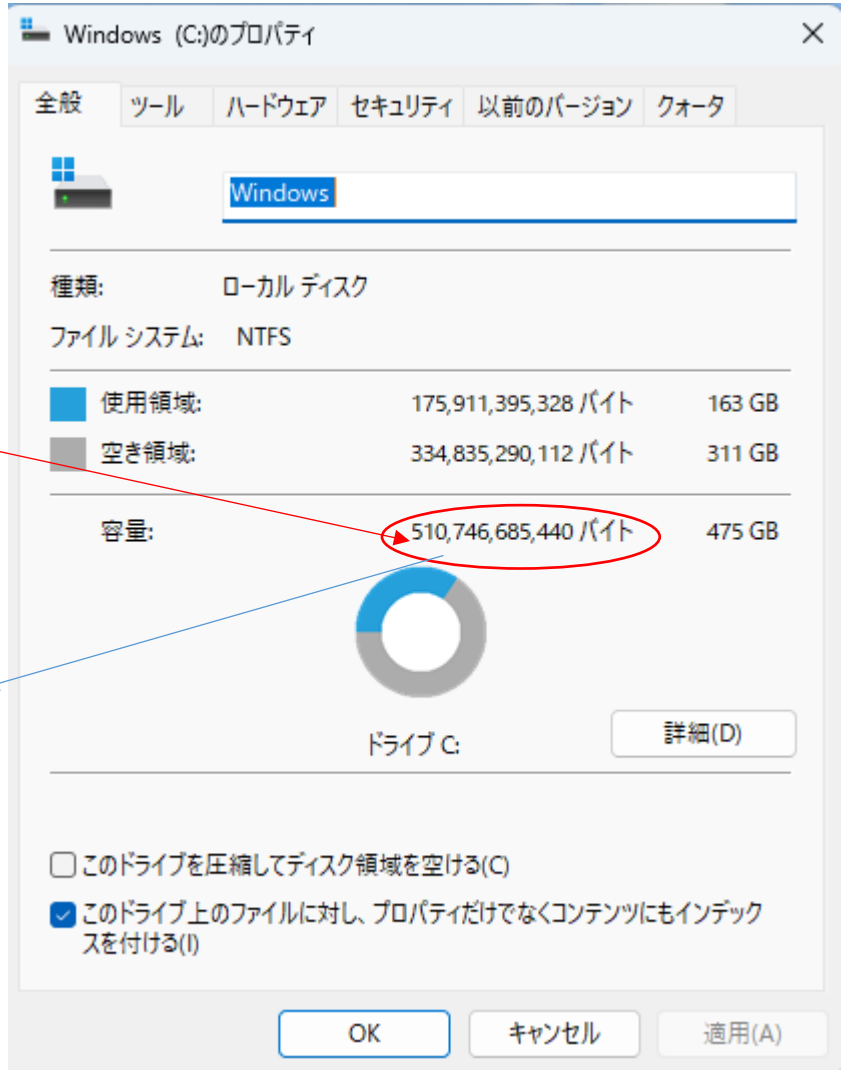
- しかしWindows (C:)のプロパティでは右図の通り

例：Windows 11 PCのディスク容量の表示

- 購入時の仕様書では：
  512GB SSD

1GBの誤差

OSが見つけられる
「認識容量」



演習

- 上記Windows上に示してある510,746,685,440Bが475GiBであることを（PC上の電卓などを利用して）確かめなさい
- これは約何GBか答えなさい

文字の表現

文字の表現

- 文字はビット列で表現される。これを符号化という
- このビット列を文字コードと呼ぶ
- 文字とビット列の対応関係を文字コード体系とよぶ。単に文字コードとよぶことも多い。以降もそう呼ぶ
- 文字コードの種類
  - 1バイト系：半角文字用。ASCIIコードなど
  - 2バイト系：全角文字用。JIS, EUC-JP, Shift_JIS（日本語）、GB, Big5（中国語）など
  - マルチバイト系：Unicode。世界で使われる全ての文字を共通の文字コードにしようという考えで作られている。
    - UTF-8：8ビットを、意味をなす最小単位としている。これを符号単位と呼ぶ。1文字を1個～6個の符号単位で符号化
    - UTF-16：16ビットを符号単位とする。1文字を1個または2個の符号単位で符号化

ASCII（アスキー）

- American Standard Code for Information Interchangeの略
- アルファベットや数字、記号を扱うことのできるもっとも基本的な文字コード
- アルファベット26種類x2（大文字と小文字）、数字10 種類、その他の記号数十種類の、計100 種類ほどの文字が扱えれば十分なので、1 文字を7 ビットで表して2^7 = 128 種類の文字を表現できるように作られている
- コンピュータは8 ビット単位で扱うものがほとんどであるため、最上位に0 を付加して8 ビットのコードとみなすことが多い

ASCIIの構成

コード範囲：16進（10進表記）	内容
00-1F (0-31)	制御文字。たとえば、改行（LF）や水平タブ（HT）など
20-7E (32-126)	印字可能文字。a, b, cなど一般的な意味での文字。空白文字(SP)を含む
7F （127）	制御文字。抹消（DEL）

制御文字（control character）とは、ディスプレイ・プリンター・通信装置などに対して、特別な動作（制御）をさせるために使う文字である
DEL: カーソルのすぐ右の文字を削除するのに使われる

ASCIIの印字可能文字

下位4ビット（16進表記）

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
2	SP	!	“	#	$	%	&	‘	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	

上位3ビット（16進表記）

ASCIIの印字可能文字

下位4ビット（16進表記）

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
2	SP	!	“	#	$	%	&	‘	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	

上位3ビット（16進表記）

たとえば、表から文字Mは$(4D)_{16}$に対応していることがわかる。つまり、$(1001101)_2$というビット列で符号化されている。これを10進数に直すと77となるので、「Mはコンピュータの中では、10進数の77で表されている」とも言える

演習03-1

1. アルファベット大文字（A-Z）の範囲を10進数で答えなさい
2. ASCII表からa, mの16進数が61, 6Dで、Aが41であることがわかる。6D-61+41を計算しその値でASCIIの表から該当する文字を答えなさい

   注意：手計算なので、補数とかコンピュータのやり方に従う必要はない。つまり、普段の10進数の計算と同じやり方でよい

manaba: 10分

演習03-2

演習03-1から教員の解説も含めヒントを得て、簡単なプログラムを解読する（実行結果を答える）問題

manaba:8分

UTF-8

- ASCII: 0x00～0x7F　（C言語などでは0xで16進数を表す）
- 各国アルファベット: 0xc080～0xdfbf
- 東アジア・インド系の諸文字（全角、半角系）、絵文字など: 0xe08080～0xefbfbf
- 上記からわかること:
  - （UTF-8の）ASCII文字はASCII文字コードと同じ
  - 東アジア各国の文字などは3バイト（4bit x 6個=24bit=3byte）で符号化　（ここでいうASCIIの最上位は8ビット中の最上位）
  - アスキーか日本語文字かは最上位ビットが0か1で区別できる。これにより、1バイトか3バイトで文字を構成するかが決まり、たとえば英文字日本語が混在する文字列からそれぞれの文字を切り出すことが可能

Question

- なぜアスキーか日本語文字かは最上位ビットが0か1で区別できるか？

日本語の場合（詳細）

文字コード	文字種類	1文字のバイト数
EUC-JP	ASCII	1
	半角カナ	1
	全角日本語	2
UTF-8	ASCII	1
	半角カナ	3
	全角日本語	3