Wykład 9: Przydział i zarządzanie pamięcią.

dr inż. Andrzej Stafiniak

Wrocław 2023





architektura von Neumanna - vs. - architektura harwardzka?



➤ Pamięć komputerowa - ?



- Pamięć komputerową stanowią "moduły komputera, służące do przechowywania danych i programów", czyli informacji. [wikipedia.org]
- Ogólny podział pamięci komputerowej: ?



- Pamięć komputerową stanowią "moduły komputera, służące do przechowywania danych i programów", czyli informacji. [wikipedia.org]
- Ogólny podział pamięci komputerowej:

Pamięć **wewnętrzna** (podstawiowa, primary storage) – bezpośrednio dostępna dla procesora. Znajdują się w niej informacje, które są potrzebne w trakcie pracy komputera dla bieżąco wykonywanych programów. Jest to najczęściej pamięć ulotna RAM (*Random Access Memory*):

- pamięć operacyjna,
- rejestry procesora,
- pamięć podręczna procesora,
 ale również pamięć nieulotna ROM (Read Only Memory):
- BIOS.

Pamięć **zewnętrzna** (masowa, secondary storage) – procesor nie ma do niej bezpośredniego dostępu. Zaliczamy do tej grupy zarówno nośniki pamięci w formie wewnętrznych i zewnętrzach dysków twardych (HDD, SDD) oraz przenośne nośniki danych takie jak półprzewodnikowe pamięci USB (pendrive'y), płyty CD, DVD, Blu-ray.

W celu obsługi informacji zawartych na tej pamięci musi być ona najpierw wczytana do pamięci operacyjnej.



- Pamięć komputerową stanowią "moduły komputera, służące do przechowywania danych i programów", czyli informacji. [wikipedia.org]
- Ogólny podział pamięci komputerowej:

Pamięć **wewnętrzna** (podstawiowa, primary storage) – bezpośrednio dostępna dla procesora. Znajdują się w niej informacje, które są potrzebne w trakcie pracy komputera dla bieżąco wykonywanych programów. Jest to najczęściej pamięć ulotna RAM (*Random Access Memory*):

- pamięć operacyjna,
- rejestry procesora,
- pamięć podręczna procesora,
 ale również pamięć nieulotna ROM (Read Only Memory):
- BIOS.

Pamięć **zewnętrzna** (masowa, secondary storage) – procesor nie ma do niej bezpośredniego dostępu. Zaliczamy do tej grupy zarówno nośniki pamięci w formie wewnętrznych i zewnętrzach dysków twardych (HDD, SDD) oraz przenośne nośniki danych takie jak półprzewodnikowe pamięci USB (pendrive'y), płyty CD, DVD, Blu-ray.

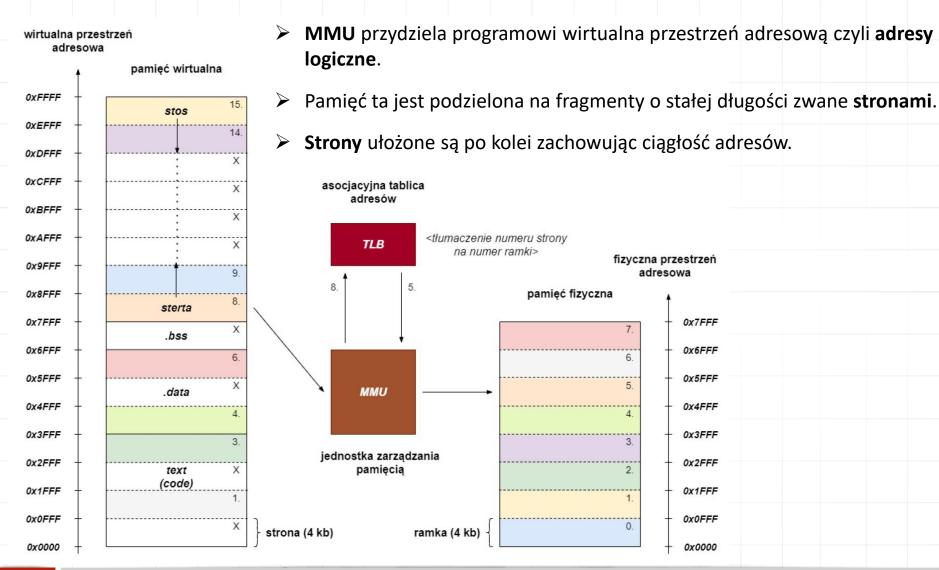
W celu obsługi informacji zawartych na tej pamięci musi być ona najpierw wczytana do pamięci operacyjnej.



- W wyniku uruchomienia programu na komputerze, do jego obsługi zostaje przydzielona pamięć,
- > jednak **nie** jest to **pamięć fizyczna** (komórki pamięci operacyjnej),
- tylko pamięć wirtualna.
- Zarządzanie pamięcią wirtualną i fizyczną jest realizowane przez jednostkę zarzadzania pamiecią MMU (ang. Memory Management Unit) z wykorzystaniem asocjacyjnej tablicy adresów TLB (ang. Translation Lookaside Buffer).

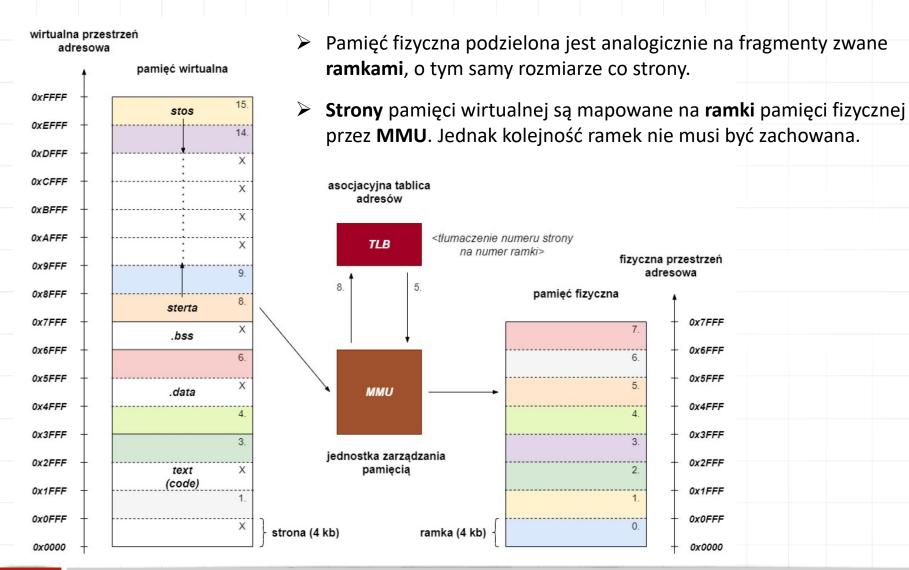


Translacja adresów logicznych na fizyczne



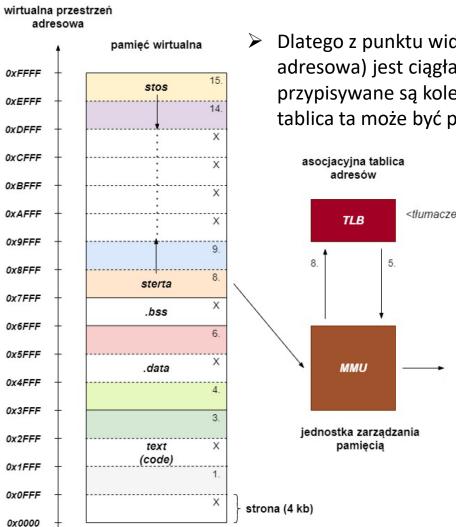


Translacja adresów logicznych na fizyczne





Translacja adresów logicznych na fizyczne

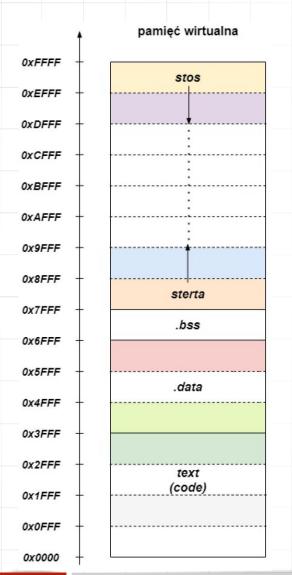


Dlatego z punktu widzenia programu pamięć (wirtualna przestrzeń adresowa) jest ciągła (przykładem są elementy tablicy, którym przypisywane są kolejne adresy komórek pamięci), ale w rzeczywistości tablica ta może być podzielona na różne sektory pamięci operacyjnej RAM.

strony n	a numer	ramki>
	strony n	strony na numer

Nr strony	Zakres adresów wirtualnych	Nr ramki	Zakres adresów fizycznych	
0	$0\mathrm{x}0000-0\mathrm{x}0\mathrm{FFF}$	_	_	
1	$0\mathrm{x}1000-0\mathrm{x}1\mathrm{FFF}$	6	0x6000 - 0x6FFF	
2	$0\mathrm{x}2000-0\mathrm{x}2\mathrm{FFF}$	_	_	
3	0x3000 - 0x3FFF	2	0x2000-0x2FFF	
4	$0\mathrm{x}4000-0\mathrm{x}4\mathrm{FFF}$	4	$0\mathrm{x}4000-0\mathrm{x}4\mathrm{FFF}$	
5	$0\mathrm{x}5000-0\mathrm{x}5\mathrm{FFF}$	_	_	
6	0x6000 - 0x6FFF	7	0x7000 - 0x7FFF	
7	0x $7000 - 0$ x 7 FFF	_	_	
8	0x8000 - 0x8FFF	5	0x5000-0x5FFF	
9	0x9000 - 0x9FFF	0	0x0000-0x0FFF	
10	0xA000 - 0xAFFF	_	_	
11	0xB000 - 0xBFFF	_	_	
12	0xC000 - 0xCFFF	_	_	
13	0xD000 - 0xDFFF	_	_	
14	0xE000 - 0xEFFF	3	0x3000 - 0x3FFF	
15	0xF000 - 0xFFFF	1	0x1000 - 0x1FFF	

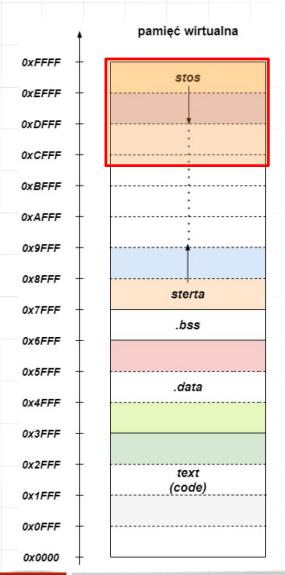




W zależności od przeznaczenia, pamięć wirtualna podzielona jest na **segmenty** :

- > stos (stack)
- > sterta (heap)
- .bss (block started by symbol).
- > .data
- text (code)

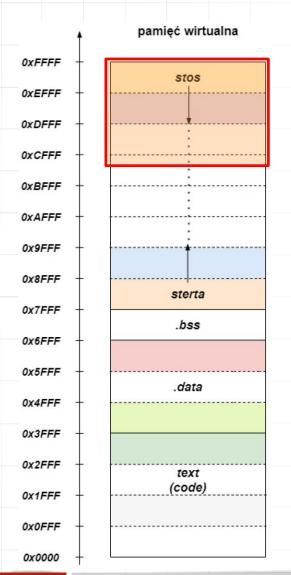




Segment **stos** (*stack*):

- ➤ Jest to liniowa struktura danych w realizacji bufora LIFO, w której elementy dodawane i usuwane sa z wierzchołka stosu (Last In, First Out ostatni element na wejściu jest pierwszym elementem na wyjściu)
- Najczęściej **stos** ulokowany jest w górnym zakresie wirtualnej przestrzeni adresowej.
- Rozmiar stosu znany jest na etapie kompilacji, a dane w nim alokowane są zwalniane automatycznie.
- ➤ Kolejne elementy dodawane na stos powodują jego rozrost w kierunku niższych adresów.

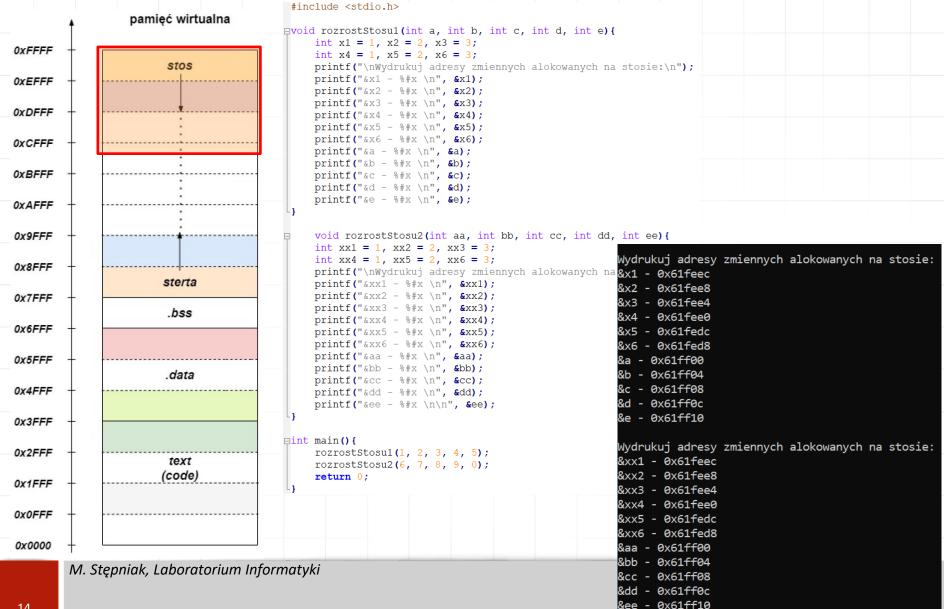


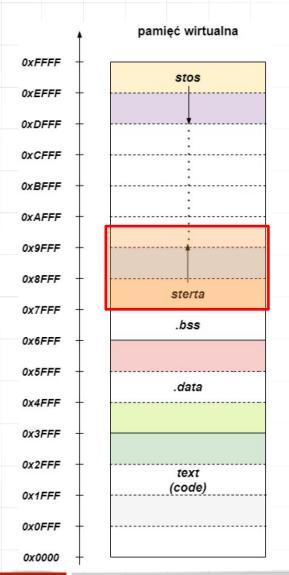


Segment **stos** (*stack*):

- na stosie przydzielana jest pamięć dla zmiennych lokalnych (automatycznych, zmiennych alokowanych statycznie - ale nie zmiennych statycznych! czas życia zmiennych lokalnych automatycznych - od definicji do końca bloku instrukcji, a zmiennych lokalnych statycznych (static) od deklaracji do końca programu)
- > na stosie umieszcza się wartości zwracane przez funkcje oraz adresy powrotu z funkcji,
- na stosie umieszcza kopie argumentów wywołania funkcji.

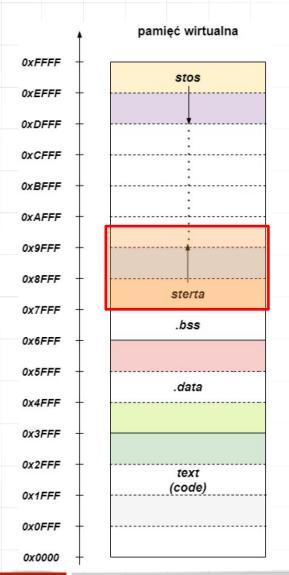






Segment **sterta** (*heap*):

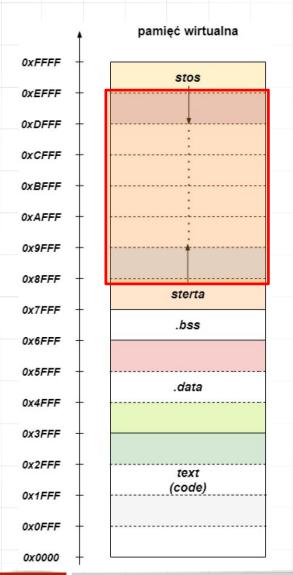
- jest to największy obszar pamięci dostępny dla programu,
- służy do przechowywania zmiennych alokowanych dynamicznie (rozmiar pamięci alokowanej na stercie nie jest znany w czasie kompilacji programu),
- przechowuje niepowiązane ze sobą elementy (w sposób nieuporządkowany, możliwa fragmentacja pamięci), do których można odwołać się z każdego miejsca w programie, jeżeli zna się ich adresy,



Segment **sterta** (*heap*):

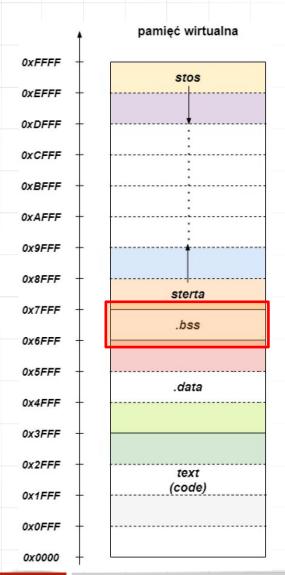
- do alokacji pamięci na stercie wykorzystuje się wskaźniki oraz dedykowane funkcje/operatory:
- funkcje (język C) malloc(), calloc(),
 realloc(),
- operatory (język C++) new, new[],
- pamięć zaalokowana na stercie nie jest zwalniana w sposób automatyczny, język C/C++ nie zapewniają mechanizmu automatycznego odśmiecania (ang. garbage collector) znanego z języków wysokiego poziomu,
- musimy pamiętać o zwolnieniu zalakowanej pamięci, funkcja free() (język C), operator delete, delete [].





Segment **sterta** (*heap*):

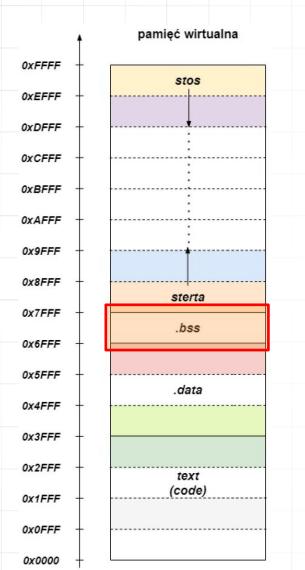
- sterta i stos współdzieli dostępny (na rozrost) obszar pamięci ze sobą,
- najczęściej adresy stosu i sterty "rosną" w przeciwnych kierunkach. Stos "rośnie" w kierunku niższych adresów, a sterta w kierunku wyższych adresów, więc istnieje możliwość nałożenia się przestrzeni adresowych obu segmentów.



Segment .bss (block started by symbol):

- przechowuje niezainicjalizowane zmienne globalne i statyczne,
- zmienne globalne i statyczne jawnie zainicjalizowane wartością 0.

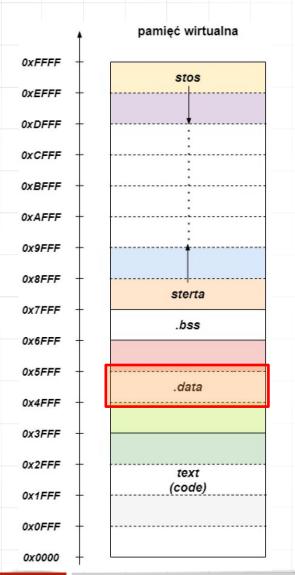
W języku C/C++ **globalne** i **statyczne** zmienne są automatycznie inicjalizowane wartością 0 przed wywołaniem funkcji main().



Segment .bss (segment danych niezainicjalizowanych):

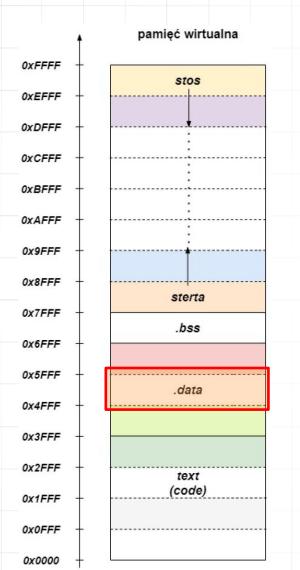
zmienne alokowane w segmencie .bss:

```
1 // Niezainicjalizowana zmienna globalna
 3 // Zmienna globalna zainicjalizowana do 0
 | int y = 0;
5 // Niezainicjalizowana statyczna zmienna globalna
 static short z;
7 // Statyczna zmienna globalna zainicjalizowana do 0
 static unsigned int w = 0;
// Zmienne umieszczone w anonimowej przestrzeni nazw
     sa rownowazne statycznym zmiennym globalnym
10 namespace {
     // Zmienna niezainicjalizowana
     double a:
     // Zmienna zainicjalizowana do 0
     long b = 0;
17 int main() {
     // Niezainicjalizowana statyczna zmienna lokalna
     static long long c;
     // Statyczna zmienna lokalna zainicjalizowana do
     static unsigned short d = 0;
     return 0;
```



Segment .data (segment danych zainicjalizowanych):

- przechowuje globalne i statyczne zmienne jawnie zainicjalizowane przez programistę (wartością inną od 0).
- W segmencie .data można wydzielić obszar pamięci,
 w którym alokowane są zmienne tylko do odczytu –
 .rodata (read-only data), gdzie alokowane są stałe.
- Rozmiar segmentu .data nie zmienia się w trakcie działania programu.

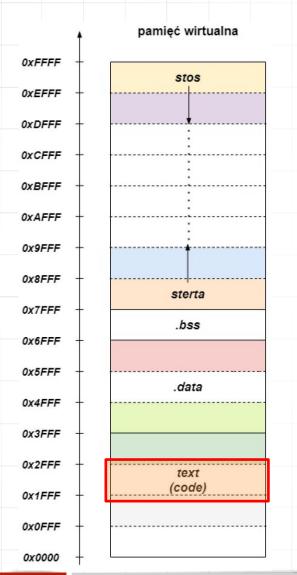


Segment .data (segment danych zainicjalizowanych):

> zmienne alokowane w segmencie .data:

```
// Zainicjalizowana zmienna globalna
_{2} float x = 3.5;
3 // Zainicjalizowane statyczne zmienne globalne
 static short y = 100;
 namespace {
      double z = 5.5;
 // Globalna stala (.rodata)
9 const int a = 111;
10 // Statyczna globalna stala (.rodata)
static const float b = 4.0;
13 int main() {
     // Zainicjalizowana statyczna zmienna lokalna
      static long w = 666;
     // Lokalna stala (.rodata)
     const unsigned short c = 1;
     return 0;
19 }
```





Segment **text** (**code**):

- > służy do przechowania skompilowanego kodu maszynowego programu,
- ma stały rozmiar,
- w większości przypadków segment text jest obszarem pamięci tylko do odczytu, aby zapobiec przed nieintencjonalna modyfikacja jego zawartości w trakcie działania programu.
- W segmencie **text** położone są również adresy funkcji zdeklarowanych w kodzie programu.



Segmenty pamięci są dodatkowo rozdzielone przerwami o losowej szerokości. Ma to na celu utrudnienie manipulacji z zewnątrz w wykonywanyn kodzie, dzięki wprowadzeniu elementu losowości w rozmieszczeniu segmentów pamięci.

Typowy układ pamięci:

wirtualna przestrzeń adresowa 0x7FFFFFF dane systemowe przerwa o losowej szerokości stos obszar dostępny dla **RAM** stosu i sterty sterta przerwa o losowej szerokości .bss .data .rodata przerwa o losowei szerokości pamieć tylko do odczytu text (code)



0x00000000

Dynamiczna alokacja pamięci -

- inaczej przydział pamięci podczas działania programu. Polega na alokowaniu pamięci na **stercie**, a rozmiar przydzielanej pamięci nie musi być znany czasie kompilacji programu.

Zastosowanie dynamicznej alokacji pamięci:

- gdy rozmiar wymaganej pamięci nie jest znany w czasie kompilacji, np. ilość elementów tablicy jest podawana przez użytkownika,
- gdy struktury danych takiej jak tablice mogą zwiększać swój rozmiar w trakcie działania programu (optymalizacja),
- w celu implementacji złożonych struktury danych takich, jak np. listy
- w celu wydłużenia czasu życia zmiennych z bloku funkcji(dla zmiennych lokalnych), ale nie na czas działania programu (zmienne globalne czy statyczne), np. gdy zmienna zdefiniowana wewnątrz funkcji ma być dostępna dla innych funkcji.



W języku C realizacja przydziału pamięci podczas działania programu odbywa się za pomocą trzech funkcji zdeklarowanych w <stdlib.h>:

Zwolnienie pamięci przydzielonej tymi funkcjami może się odbyć wyłącznie za pomocą funkcji free().



Nagłówek funkcji:

Typ zwracany - wskaźnik na początek przydzielonej pamięci. Typ void * może przechowywać adres zmiennej dowolnego typu i może być rzutowana na konkretny typ wskaźnikowy.

Wielkość alokowanej pamięci w bajtach. Typ size_t (alias dla unsigned int) - stosowany do reprezentowania rozmiarów i liczb.

Funkcja malloc () rezerwuje na stercie niezainicjalizowaną pamięć o rozmiarze size. W przypadku poprawnego wywołania zwracany jest wskaźnik na początek zaalokowanego bloku pamięci. W innym wypadku zwracana jest wskaźnik zerowy - NULL.



Nagłówek funkcji:

void * calloc(size_t num , size_t size)

wskaźnik na początek przydzielonej pamięci

Ilość **num** elementów o rozmiarze **size** w bajtach .

Funkcja **calloc()** rezerwuje na **stercie** pamięć, oraz inicjalizuje zerami dla **num** elementów o rozmiarze **size**. W przypadku poprawnego wywołania zwracany jest **wskaźnik** na początek zaalokowanego bloku pamięci. W innym wypadku zwracana jest wskaźnik zerowy - NULL.



Nagłówek funkcji:

```
void * realloc(void * ptr , size_t size)
wskaźnik na początek
przydzielonej pamięci
nowy rozmiar
przydzielanej pamięci
```

Wskaźnik na wcześniej przydzieloną pamięć za pomocą funkcji malloc(), calloc(), realloc(), jeżeli wartość wskaźnika NULL - to alokuje pamięć o rozmiarze size (jak malloc())

Funkcja **realloc** () alokuje segment pamięci o rozmiarze size w obszarze wskazywanym przez wskaźnik ptr, który został wcześniej zaalokowany i nie został zwolniony z użyciem funkcji free (). W ten sposób możliwe jest rozszerzanie bądź zmniejszanie zaalokowanego obszaru pamięci (optymalizacja). Należy mieć na uwadze, ze funkcja realloc (), w pierwszej kolejności zwalnia pamięć wskazywaną przez wskaźnik ptr.



Nagłówek funkcji:

Wskaźnik na wcześniej przydzieloną pamięć za pomocą jednej z funkcji malloc(), calloc(), realloc().

Funkcja **free** () zwalnia pamięć zaalokowana za pomocą jednej z funkcji malloc(), calloc() albo realloc(), którą wskazuje ptr (jeśli ptr jest wskaźnikiem na pamięć zaalokowaną w inny sposób, to wynik działania funkcji jest niezdefiniowany. Jeżeli ptr ma wartość NULL to funkcja nie robi nic.

Funkcja free () nie zmienia wartości wskaźnika (nie ustawia wartości na NULL) ani nie zeruje zalakowanej wcześniej pamięci !!!



```
#include <stdio.h>
  #include <stdlib.h> /* realloc, calloc, malloc, free, exit, NULL */
  int main ()
                           //liczba i l-licznik
      int liczba=0, l = 1;
      int * tabInt = NULL;
      int * aktTabInt = NULL;
      /*
      printf ("Wprowadz rozmiar tablicy dynamicznej");
      scanf ("%d", &liczba);
      tabInt = (int*) malloc(liczba*sizeof(int));
      tabInt = (int*) calloc(liczba, sizeof(int));
      * /
      do{
          printf ("Wprowadz liczbe calkowita (666 - stop): ");
          scanf ("%d", &liczba);
          if (liczba == 666)
              break;
          aktTabInt = (int*) realloc(tabInt, l*sizeof(int));
          if (aktTabInt!=NULL) {
              tabInt=aktTabInt;
              tabInt[1-1]=liczba;
          else{
              free (tabInt);
              printf("Blad utworzenia lub optymalizacji tablicy");
              exit(EXIT FAILURE);
                                                             Wprowadz liczbe calkowita (666 - stop): 1
          ++1;
                                                             Wprowadz liczbe calkowita (666 - stop): 2
      }while (1);
                                                             Wprowadz liczbe calkowita (666 - stop): 3
                                                             Wprowadz liczbe calkowita (666 - stop): 4
      printf ("Wprowadzone liczby to: ");
      for (int i=0; i<1-1;++i)
                                                             Wprowadz liczbe calkowita (666 - stop): 5
          printf ("%d, ",aktTabInt[i]);
                                                              Wprowadz liczbe calkowita (666 - stop): 6
      free(aktTabInt);
                                                             Wprowadz liczbe calkowita (666 - stop): 666
      return 0;
30
                                                             Wprowadzone liczby to: 1, 2, 3, 4, 5, 6,
```

```
#include <stdlib.h> /* realloc, calloc, malloc, free, exit, NULL */
int main ()
                         //liczba i l-licznik
   int liczba=0, 1 = 1;
   int * tabInt = NULL;
   int * aktTabInt = NULL;
   /*
   printf ("Wprowadz rozmiar tablicy dynamicznej");
   scanf ("%d", &liczba);
   tabInt = (int*) malloc(liczba*sizeof(int));
   tabInt = (int*) calloc(liczba, sizeof(int));
   * /
                                                              W języku C/C++ nie ma możliwości
   do{
                                                              sprawdzenia rozmiaru dynamicznej
       printf ("Wprowadz liczbe calkowita (666 - stop): ");
       scanf ("%d", &liczba);
                                                              tablicy za pomocą operatora sizeof.
       if (liczba == 666)
          break:
       aktTabInt = (int*) realloc(tabInt, l*sizeof(int));
                                                              Wartość tą musimy zapamiętać oraz
       if (aktTabInt!=NULL) {
                                                              ewentualnie przekazać.
           tabInt=aktTabInt;
          tabInt[1-1]=liczba;
       else{
           free (tabInt);
          printf("Blad utworzenia lub optymalizacji tablicy");
           exit(EXIT FAILURE);
                                                         Wprowadz liczbe calkowita (666 - stop): 1
       ++1;
                                                         Wprowadz liczbe calkowita (666 - stop): 2
   }while (1);
                                                         Wprowadz liczbe calkowita (666 - stop): 3
                                                         Wprowadz liczbe calkowita (666 - stop): 4
   printf ("Wprowadzone liczby to: ");
   for (int i=0;i<1-1;++i)</pre>
                                                         Wprowadz liczbe calkowita (666 - stop): 5
       printf ("%d, ",aktTabInt[i]);
                                                         Wprowadz liczbe calkowita (666 - stop): 6
   free(aktTabInt);
                                                         Wprowadz liczbe calkowita (666 - stop): 666
   return 0;
                                                         Wprowadzone liczby to: 1, 2, 3, 4, 5, 6,
```

#include <stdio.h>

Dynamiczna alokacja pamięci – język C++ <iostrem>

Nagłówek funkcji:

```
void * operator new(std::size_t size);
void operator delete(void * ptr);
```

Domyślne operatory są specjalnymi składnikami biblioteki standardowej, a ich deklaracja znajduje się w globalnej przestrzeni nazw. Można dołączyć nagłówek <new> lecz nie jest to konieczne.

Operator new alokuje dynamicznie pamięć na stercie dla zmiennej/obiektu o rozmiarze size bajtów. Jednak rozmiar nie jest to podany jawnie tylko po przez wskazanie typu. W przypadku powodzenia zwracany jest wskaźnik na początek zaalokowanego bloku pamięci. W innym przypadku wyrzucany jest wyjątek bad_alloc (przy odpowiedniej konstrukcji new (std :: nothrow) zwracany jest nullptr).

Operator **delete** zwalnia pamięć wskazywaną przez wskaźnik **ptr**, która została przydzielona za pomocą operatora **new**. Jeżeli **ptr** ma wartosc **nullptr**, to operator **delete** nie robi nic, w każdej innej sytuacji zachowanie jest niezdefiniowane.



Dynamiczna alokacja pamięci – język C++ <iostrem>

Nagłówek funkcji:

```
void * operator new[](std::size_t size);
void operator delete[](void * ptr);
```

Oba poprzednie operatory występują również w wersji do dynamicznej alokacji pamięci dla tablic obiektów, new[], delete[]. Ważne aby operatorem delete[] zwalniać pamięć przydzieloną za pomocą operatora new[].

Operatory new, new[], delete, delete[] mogą być przeciążane (przeciążanie/ przeładowanie operatorów pozwala na nadanie im nowych funkcji)



Dynamiczna alokacja pamięci – język C++ <iostrem>

```
#include <iostream>
int main(){
    unsigned int tabSize;
    std::cout << "Podaj rozmiar tablicy tab[n][n] = ";</pre>
    std::cin >> tabSize;
    // Deklaracja dynamicznej tablicy dwuwymiarowej
    int ** tab = new int * [tabSize];
    for (unsigned int i = 0; i<tabSize; ++i)</pre>
        tab[i] = new int [tabSize];
    // Inicjalizacja tablicy
    std::cout << "Wprowadz kolejne elementy tablicy tab[i][j] rozdzielone spacja:";</pre>
    std::cout << std::endl;</pre>
    for ( unsigned int i = 0; i < tabSize; ++i){</pre>
        for (unsigned int j = 0; j < tabSize; ++j){}
            std :: cin >> tab[i][i];
    // Wyswietlenie zawartosci tablicy
    std::cout << "Wprowadzone wartosci do tab[i][j] to:" << std::endl;</pre>
    for ( unsigned int i = 0; i < tabSize; ++i){</pre>
        for (unsigned int j = 0; j < tabSize; ++j){
            std :: cout << tab[i][j] << " ";
                                         Podaj rozmiar tablicy tab[n][n] = 3
        std::cout << std::endl;</pre>
                                         Wprowadz kolejne elementy tablicy tab[i][j] rozdzielone spacja:
                                         1 2 3 4 5 6 7 8 9
    // Zwolnienie pamieci
                                         Wprowadzone wartosci do tab[i][j] to:
    for (int i = 0; i<tabSize; i++)</pre>
        delete [] tab[i];
                                                 6
    delete [] tab ;
```