

Planowanie przydziału procesora

Ogólna koncepcja planowania

- Tryb decyzji — określa moment czasu, w którym oceniane i porównywane są priorytety procesów i dokonywany jest wybór procesu do wykonania.
- Funkcja priorytetu — funkcja zwracająca aktualny priorytet procesu na podstawie parametrów procesu i stanu systemu.
- Reguła arbitrażu — reguła rozwiązywania konfliktów pomiędzy procesami o tym samym priorytecie.

Komponenty jądra w planowaniu

- Planista krótkoterminowy (ang. CPU scheduler) — wyznacza wartość priorytetu procesów gotowych i wybiera proces (o najwyższym priorytecie) do wykonania.
- Ekspedytor (ang. dispatcher) — realizuje przekazanie sterowanie do procesu wybranego przez planistę (jest to najczęściej ostatni krok w przełączeniu kontekstu).

Tryb decyzji

- Schemat niewyłączeniowy (ang. nonpreemptive) — proces po uzyskaniu dostępu do procesora wykonywany jest do momentu zakończenia lub zgłoszenia żądania obsługi do systemu.
- Schemat wyłączeniowy (ang. preemptive) — proces może zostać zatrzymany i umieszczony w kolejce procesów gotowych, a procesor zostaje przydzielony procesowi o wyższym (lub równym) priorytecie.

Podejmowanie decyzji o wyłączeniu

- Utworzenie i przyjęcie nowego procesu.
- Obudzenie procesu w wyniku otrzymania komunikatu, sygnału gotowości urządzenia (przerwanie) lub sygnału wynikającego z synchronizacji.
- Upływanie kwantu czasu odmierzanego przez czasomierz.
- Wzrost priorytetu innego procesu w stanie *gotowy* powyżej priorytetu procesu wykonywanego — możliwe w systemie z dynamicznym mechanizmem zmiany priorytetów.

Wyłączanie selektywne

- Z każdym procesem wiąże się parę bitów (u_p, v_p) o następującej interpretacji:

$$u_p = \begin{cases} 1, & \text{jeśli } p \text{ może wyłączyć inny proces} \\ 0, & \text{w przeciwnym przypadku} \end{cases}$$

$$v_p = \begin{cases} 1, & \text{jeśli } p \text{ może zostać wyłączony} \\ 0, & \text{w przeciwnym przypadku} \end{cases}$$

Uogólnione wywłaszczanie selektywne (1)

- ✿ Z każdym procesem wiąże się parę priorytetów (Π_p, Φ_p) , w której Π_p oznacza priorytet procesu w stanie gotowości, a Φ_p oznacza priorytet procesu wykonywanego.
- ✿ Proces q może wywłaszczyć proces p , gdy $\Pi_q > \Phi_p$

Uogólnione wywłaszczanie selektywne (2)

- ✿ Uniwersalny mechanizm wywłaszczania selektywnego można by uzyskać, implementując macierz bitów, określających prawo wywłaszczania.
- ✿ Ustawiony bit na pozycji (p, q) oznaczałby prawo wywłaszczania procesu q przez proces p .

Klasyfikacja SO ze względu na sposób przetwarzania

- ✿ Systemy przetwarzania bezpośredniego (ang. on-line processing systems) — systemy interakcyjne
 - występuje bezpośrednia interakcja pomiędzy użytkownikiem a systemem
 - wykonywanie zadania użytkownika rozpoczyna się zaraz po przedłożeniu
- ✿ Systemy przetwarzania pośredniego (ang. off-line processing systems) — systemy wsadowe
 - występuje znacząca zwłoka czasowa między przedłożeniem a rozpoczęciem wykonywania zadania
 - niemożliwa jest ingerencja użytkownika w wykonywanie zadania

Pytania

- ✿ Jaki tryb decyzji obowiązuje w systemach przetwarzania bezpośredniego, a jaki w systemach wsadowych?
- ✿ Jak wygląda macierz praw wywłaszczania w systemach przetwarzania bezpośredniego, a jak w systemach wsadowych?

Funkcja priorytetu

- ✿ Argumentami funkcji priorytetu są parametry procesu oraz stanu systemu.
- ✿ Priorytet procesu jest wartością funkcji priorytetu dla bieżących wartości parametrów danego procesu i aktualnego stanu systemu.

Parametry funkcji priorytetu (1)

- ✿ wymagania odnośnie wielkości przestrzeni adresowej pamięci,
- ✿ czas oczekiwania — czas spędzony w kolejce procesów gotowych (czas spędzony w stanie gotowości)
- ✿ czas obsługi — czas, przez który proces był wykonywany (wykorzystywał procesor) od momentu przyjęcia do systemu

Parametry funkcji priorytetu (2)

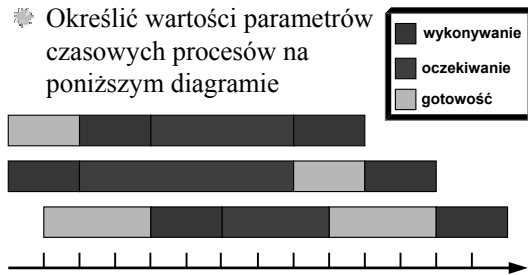
- ☼ rzeczywisty czas przebywania w systemie — czas spędzony w systemie od momentu przyjęcia (czas obsługi + czas oczekiwania + czas realizacji żądań zasobowych),
- ☼ priorytet zewnętrzny — składowa priorytetu, która pozwala wyróżnić procesy ze względu na klasy użytkowników.

Parametry funkcji priorytetu (3)

- ☼ czasowa linia krytyczna — określa czas po którym wartość wyników spada (nawet do zera, np. przy przewidywaniu pogody),
- ☼ obciążenie systemu — liczba procesów przebywających w systemie i ubiegających się (potencjalnie) o przydział procesora.

Zadanie

- ☼ Określić wartości parametrów czasowych procesów na poniższym diagramie



Reguła arbitrażu

- ☼ losowo — możliwe w przypadku, gdy liczba procesów o tym samym priorytecie jest niska,
- ☼ cyklicznie — cykliczny przydział procesora kolejnym procesom,
- ☼ w kolejności FIFO — w kolejności przyjmowania procesów do systemu.

Kryteria oceny algorytmów planowania (1)

- ☼ Efektywność z punktu widzenia systemu
 - wykorzystanie procesora (processor utilization) — procent czasu, przez który procesor jest zajęty pracą,
 - przepustowość (throughput) — liczba procesów kończonych w jednostce czasu.
- ☼ Inne aspekty z punktu widzenia systemu
 - sprawiedliwość (fairness) — równe traktowanie proc.,
 - respektowanie priorytetów procesów
 - równoważenie obciążenia wykorzystania zasobów

Kryteria oceny algorytmów planowania (2)

- ☼ Efektywność z punktu widzenia użytkownika
 - czas cyklu przetwarzania (turnaround time) — czas pomiędzy przedłożeniem zadania, a zakończeniem jego wykonywania (rzeczywisty czas przebywania w systemie w momencie zakończenia procesu),
 - czas odpowiedzi (response time) — czas pomiędzy przedłożeniem zadania, a uzyskaniem pierwszej odpowiedzi.
- ☼ Inne aspekty z punktu widzenia użytkownika
 - przewidywalność — realizacja przetwarzania w zbliżonym czasie niezależnie od obciążenia systemu.

Algorytmy planowania niewyłączającego

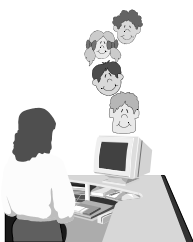
- ☼ FCFS — pierwszy zgłoszony, pierwszy obsługony,
- ☼ LCFS — ostatni zgłoszony, pierwszy obsługony,
- ☼ SJF (SJN) — najpierw najkrótsze zadanie,
- ☼ planowanie priorytetowe — bazujące na priorytecie zewnętrznym,
- ☼ planowanie przed liniami krytycznymi — zakończenie zadania przed czasową linią krytyczną lub możliwie krótko po tej linii.

Algorytmy planowania wyłączającego

- ☼ Planowanie rotacyjne — po ustalonym kwancie czasu proces wykonywany jest przerywany i trafia do kolejki procesów gotowych.
- ☼ SRT — najpierw zadanie, które ma najkrótszy czas do zakończenia,
- ☼ Planowanie wielokolejkowe — w systemie jest wiele kolejek procesów gotowych i każda z kolejek może być inaczej obsługiwana.

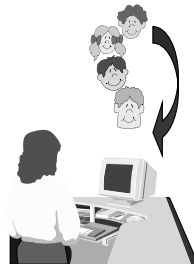
FCFS (FIFO)

- ☼ Wykonywanie procesów w kolejności zgłaszania się do systemu
- ☼ Duży rozrzut czasu oczekiwania



LCFS (LIFO)

- ☼ Wykonywanie procesów w kolejności odwrotnej do kolejności zgłaszania się do systemu
- ☼ Podejście nie stosowane w współczesnych systemach komputerowych.



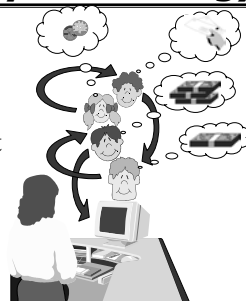
SJF (SJN)

- ☼ Wybierany jest proces, który ma najkrótszy czas obsługi.
- ☼ Daje minimalny średni czas oczekiwania



Planowanie priorytetowe (ang. priority scheduling)

- ☼ Wybierany jest proces, który ma największy priorytet zewnętrzny.



Planowanie rotacyjne (ang. Round Robin — RR)

- ☼ Po upływie ustalonego kwantu czasu proces jest **wyłączany** i trafia na koniec kolejki procesów gotowych (chyba że wcześniej zażąda operacji wejścia-wyjścia)
- ☼ Preferencja dla zadań krótkich (wydłuża się czas oczekiwania i czas cyklu przetwarzania dla zadań długich)
- ☼ Przełączanie kontekstu pochłania pewien czas!!!



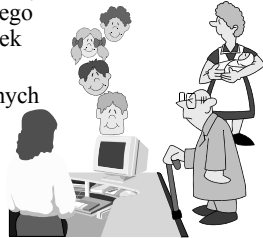
SRT

- ☼ Wybierany jest proces, który ma najkrótszą następną fazę procesora (wyłączająca wersja alg. SJF)



Planowanie wielokolejkowe

- ☼ Podział procesów na grupy (np. procesy interaktywne i procesy wsadowe) i wynikający z tego przydział do różnych kolejek procesów gotowych
- ☼ Możliwość przydziału różnych priorytetów oraz różnych algorytmów szeregowania do poszczególnych kolejek



Zadanie

- ☼ Zdefiniować funkcję priorytetu dla algorytmów
 - FCFS
 - LCFS
 - SJN
 - SRT
- jako funkcję zmiennych (parametrów procesu)
 - a — bieżący (dotychczasowy) czas obsługi
 - r — rzeczywisty czas w systemie
 - t — całkowity (do momentu zakończenia) czas obsługi

Własności algorytmów planowania

algorytm	priorytet	tryb decyzji	arbitraż
FIFO	r	niewyłączalny	losowy
LIFO	$-r$	niewyłączalny	losowy
SJN	$-t$	niewyłączalny	losowy lub chronologiczny
SRT	$a - t$	wyłączalny (przyjęcie proc.)	losowy lub chronologiczny
RR	stały	wyłączalny (kwant czasu)	cykliczny

Implementacja algorytmów planowania

- ☼ Z punktu widzenia przetwarzania użytkowego przełączanie kontekstu jest marnotrawstwem czasu procesora.
- ☼ Decyzja planisty krótkoterminowego musi zapaść w możliwie krótkim czasie.
- ☼ Struktury danych muszą dostarczyć informacji niezbędnych do dokonania szybkiego wyboru procesu o najwyższym priorytecie zgodnie z polityką planowania przydziału procesora (modelem matematycznym).

Implementacja algorytmu FCFS

- Struktura danych dla kolejki procesów gotowych → kolejka FIFO
- Umieszczenie procesu w kolejce procesów gotowych → dopisanie procesu na końcu kolejki FIFO
- Wybór procesu do wykonania → pobranie procesu z czoła kolejki FIFO
- Czy taki algorytm realizuje dokładnie założenia modelu matematycznego?

Implementacja algorytmu LCFS

- Struktura danych dla kolejki procesów gotowych → stos
- Umieszczenie procesu w kolejce procesów gotowych → odłożenie procesu na szczycie stosu
- Wybór procesu do wykonania → zdjęcie procesu ze szczytu stosu
- Czy taki algorytm realizuje dokładnie założenia modelu matematycznego?

Implementacja algorytmu SJN

- Struktura danych dla kolejki procesów gotowych → lista posortowana rosnąco wg. założonego całkowitego czasu obsługi
- Umieszczenie procesu w kolejce procesów gotowych → wstawienie procesu do listy w kolejności zgodnej z całkowitym czasem obsługi
- Wybór procesu do wykonania → zdjęcie pierwszego procesu z listy

Algorytm FCFS w ujęciu probabilistycznym

- λ — średnia częstotliwość przyjmowania procesów do systemu
- T_s — średni czas obsługi procesu
- μ — maksymalna liczba procesów obsługiwanych w jednostce czasu

$$\mu = \frac{1}{T_s}$$

Algorytm FCFS — wykorzystanie procesora

- ρ — średnie wykorzystanie procesora

$$\rho = \frac{\lambda \cdot T_s}{\mu \cdot T_s} = \frac{\lambda}{\mu}$$

- Jakie jest wykorzystanie procesora, gdy do systemu przybywa więcej procesów, niż procesor jest w stanie obsłużyć?

Algorytm FCFS — czas oczekiwania

- $T_w(p)$ — czas oczekiwania procesu p na przydział procesora
- L — liczba procesów oczekujących w momencie przyjęcia procesu p do systemu

$$T_w(p) = L \cdot T_s + \frac{1}{2} \cdot T_s = \left(L + \frac{1}{2} \right) \cdot \frac{1}{\mu}$$

Algorytm RR — dobór kwantu czasu

- ❁ Krótki kwant czasu oznacza zmniejszenie czasu cyklu przetwarzania procesów krótkich, ale zwiększa narzut czasowy związany z obsługą przerw i przełączaniem kontekstu.
- ❁ Z punktu widzenia interakcji z użytkownikiem kwant czasu powinien być trochę większy, niż czas potrzebny na typową interakcję.

Dobór kwantu czasu, a czas odpowiedzi systemu

