

Planowanie zadań

Proces jest dynamicznym obiektem wykonującym w systemie operacyjnym określony program. **Planowaniem** (*scheduling*), lub alternatywnie **szeregowaniem**, nazywamy czynność rozdzielania czasu procesora na poszczególne procesy, zapewniającą ich *quasi-równoległe wykonywanie*.

Proces jest podstawową jednostką roboczą planowaną przez system. W niektórych systemach operacyjnych nie występują procesy, natomiast mogą być tworzone wątki i wtedy one podlegają planowaniu w podobny sposób. W kontekście systemów czasu rzeczywistego często mówi się o planowaniu **zadań** nie rozróżniając, czy są one procesami (z oddzielnym przydziałem pamięci), czy wątkami (współdzielącymi przydział pamięci programu).

Typy planowania

Rozróżnia się planowanie: krótkoterminowego, średnioterminowego, i długoterminowego.

Planowanie długoterminowe podejmuje decyzje o dodaniu procesu do puli procesów aktywnych. Ta decyzja może wpływać na stopień wieloprogramowości systemu. Ograniczając ten parametr, system może spowodować sprawniejsze wykonywanie procesów aktywnych, poprzez przydzielanie im większego ułamka czasu procesora. Te zagadnienia nie będą omawiane na tym wykładzie.

Planowanie średnioterminowe podejmuje decyzje o umożliwieniu wykonywania procesu przez umieszczenie całego lub części alokacji pamięci procesu w pamięci fizycznej komputera. Tej funkcji systemu operacyjnego przyjrzymy się na wykładzie o zarządzaniu pamięcią.

Planowanie krótkoterminowe podejmuje decyzje który proces należy w danym momencie załadować na procesor i wykonywać. W tym wykładzie zajmiemy się tym właśnie typem planowania, nazywając go krótko planowaniem, lub zamiennie szeregowaniem.

Przełączanie kontekstu

Zastępowanie procesu dotychczas wykonywanego na procesorze innym procesem, wynikające z planowania, wymaga zapisania stanu obliczeń w taki sposób, aby było możliwe ich wznowienie w sposób niezauważony przez proces. Następnie, załadowanie i uruchomienie innego procesu wymaga odtworzenia stanu wszystkich rejestrów procesora, wcześniej zapamiętanych.

Ta operacja składowania a później odtwarzania stanu procesora ze wszystkimi rejestrami nazywana jest **przełączeniem kontekstu**. **Stanowi ona istotny narzut na wykonywanie procesu, albo dodatkowe obciążenie systemu, ponieważ czas poświęcony na przełączanie kontekstu jest tracony, procesor nie wykonuje w tym czasie zadanych obliczeń.** Przełączanie kontekstu nie może zatem być wykonywane zbyt często, ponieważ w takim przypadku czas na nie tracony byłby znaczący.

Szybkość przełączania kontekstu zależy też od wsparcia sprzętowego. W niektórych architekturach występuje kilka zbiorów rejestrów. Wówczas przełączenie kontekstu może polegać na przełączeniu wskaźnika bieżącego zestawu rejestrów, i nie trzeba ich zapamiętywać i odtwarzać.

Blok kontrolny procesu PCB

Blok kontrolny procesu (*Process Control Block PCB*) zawiera główne informacje o stanie procesu. Stan procesu jest zapisywany przy przełączaniu kontekstu i odtwarzany na podstawie zawartości PCB przy wznowianiu obliczeń na procesorze. PCB zawiera również informacje o procesie wykorzystywane w algorytmach planowania. W szczególności:

- identyfikator procesu,
- adres startowy,
- kontekst: stan procesu, licznik instrukcji (PC), zawartość rejestrów,
- informacje związane z planowaniem,
- informacje związane z synchronizacją,
- informacje związane z zarządzaniem pamięcią,
- informacje dotyczące zużycia czasu,
- informacje o stanie wejścia-wyjścia,
- ...

Umieszczenie przez planistę procesu w kolejce procesów wykonywanych oznacza umieszczenie PCB tego procesu na liście innych PCB w tej kolejce. Usunięcie procesu z systemu polega na wykasowaniu jego PCB i dealokacji zajmowanej pamięci.

Planista i dyspozytor

W realizacji planowania biorą udział następujące moduły systemu operacyjnego:

- **planista** (*scheduler*) — podejmuje decyzje, który z procesów gotowych powinien być następnie wykonywany na procesorze,
- **dyspozytor** (*dispatcher*), zwany czasami alternatywnie **egzekutorem** albo **ekspedytorem** — wykonuje wszystkie czynności administracyjne związane z przełączaniem kontekstu przy przenoszeniu jednych procesów do kolejki gotowych, a innych na procesor.

Moduły planisty i dyspozytora normalnie zlokalizowane są w **jądrze** (*kernel*) systemu operacyjnego. W niektórych systemach, moduł planowania może być wydzielony z jądra (zwanego wtedy mikrojądrem) i wykonywany jako oddzielny proces.

Łączny czas pracy planisty i dyspozytora jest dodatkowym narzutem administracyjnym. Zatem ich działania muszą być podejmowane na tyle rzadko, by nie wpływały nadmiernie na efektywność działania systemu.

Z drugiej strony, **z punktu widzenia quasi-równoległego wykonywania procesów wymiana procesu obliczanego na procesorze powinna następować często.**

Jest to podstawowy kompromis, który muszą rozwiązywać strategie planowania.

Okres planowania

Wykonywanie algorytmu planisty, który wybiera następny proces do wykonywania na procesorze **następuje okresowo, w momencie przerwania zegarowego**, kiedy system aktualizuje swoje informacje o czasie i timerach, i oblicza swój budżet czasowy. Typowym okresem przerwania zegarowego jest 10 milisekund (starsze systemy, Linux 2.4), albo 1 milisekunda (Linux 2.6 i nowsze systemy).

Planista jest wywoływany również po wystąpieniu zdarzeń, które mogą mieć wpływ na decyzje planistyczne, takich jak:

- utworzenie nowego procesu, który jest gotowy i powinien być planowany,
- zakończenie procesu i zwolnienie procesora,
- dobrowolne zwolnienie procesora,
- żądanie zasobu, które nie może być natychmiast spełnione (np. semafora),
- przerwanie od urządzenia wejścia/wyjścia; jeśli jakieś urządzenie zakończyło transfer, to może trzeba uruchomić proces, który oczekiwał na te dane.

Zatem system operacyjny wykonuje jednocześnie **planowanie zegarowe i zdarzeniowe**.

Wywłaszczanie

Można wyróżnić dwa rodzaje algorytmów planowania:

niewywłaszczeniowe

Proces jest usuwany z procesora wyłącznie w momencie jego zakończenia, albo gdy sam dobrowolnie poprosi o przeniesienie do kolejki gotowych (patrz np. funkcja systemowa `sched_wait`).

Takie planowanie ogranicza do minimum aktywność planisty i dyspozytora i maksymalizuje użyteczny czas procesora.

wywłaszczeniowe

Proces może być usunięty z procesora w wyniku podjęcia takiej decyzji przez planistę w którymkolwiek z momentów wymienionych wcześniej.

Przy stosowaniu planowania niewywłaszczeniowego jest możliwe, że proces będzie wykonywał się na procesorze bardzo długo (np. wiele godzin lub dni) nie dopuszczając innych procesów do wykonania.

Zagłodzenie i sprawiedliwość

Niektóre strategie planowania mogą doprowadzić do sytuacji, kiedy pewien proces, lub procesy, będą wielokrotnie pomijane i nieuruchamiane przez planistę, w zgodzie z przyjętym algorytmem planowania. Jest to możliwe jeśli stale będą się pojawiały nowe procesy, które powinny być uruchomione wcześniej.

To zjawisko zwane jest **zagłodzeniem** procesu, i jest **niepożądanym efektem** ubocznym pewnych algorytmów planowania. W związku z tym w tych algorytmach stosuje się dodatkowe zabezpieczenia, ingerujące w algorytm planowania i nie dopuszczające do zagłodzenia.

Unikanie zagłodzenia procesów jest wyrazem ogólnej zasady „sprawiedliwości” (ang. *fairness*), albo **egalitaryzmu**, obowiązującą w systemach operacyjnych ogólnego przeznaczenia (*GPOS — General Purpose Operating System*). Zasada ta mówi, że **równoprawne z punktu widzenia systemu procesy powinny otrzymać równy dostęp do zasobów systemu**. Obowiązuje ona w przydziale: procesora, pamięci, i innych zasobów.

Zauważmy, że zasadę tę można odnosić do wszystkich procesów równo, albo też do wszystkich użytkowników, to znaczy do grup procesów poszczególnych użytkowników.

Cele planowania

Można brać pod uwagę różne kryteria jakości planowania procesów, i dostosować do nich strategię planowania. Te kryteria są jednak do pewnego stopnia, lub całkowicie, ze sobą sprzeczne. Planowanie może zatem optymalizować:

wykorzystanie procesora

średni czas efektywnych obliczeń, odwrotność czasu traconego na puste cykle, oraz czasu zużytego na przełączenia kontekstu,

przepustowość

średnią liczbę procesów wykonanych na jednostkę czasu,

czas przetwarzania

średni całkowity czas przetwarzania zadania, od pojawienia się do zakończenia,

czas oczekiwania

średni sumaryczny czas spędzony przez proces w kolejce procesów gotowych.

W dodatku do powyższych kryteriów obowiązują ograniczenia: równe traktowanie procesów i niedopuszczenie do zagłodzenia.

Krótkie podsumowanie — pytania sprawdzające

Odpowiedz na poniższe pytania:

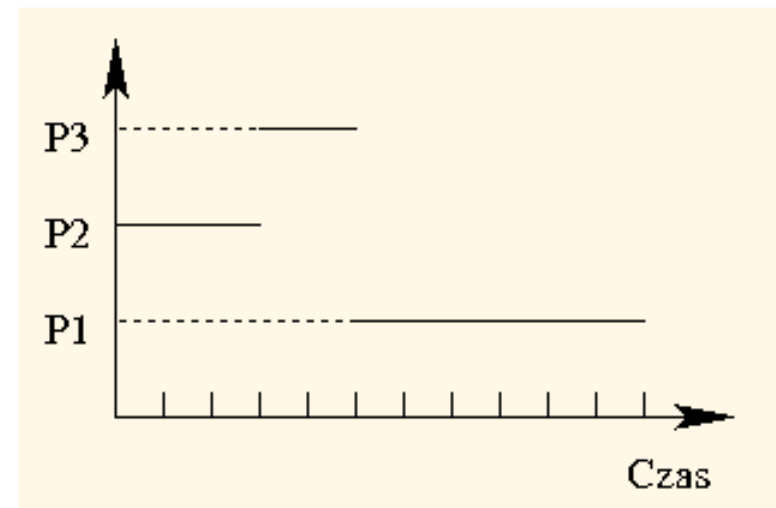
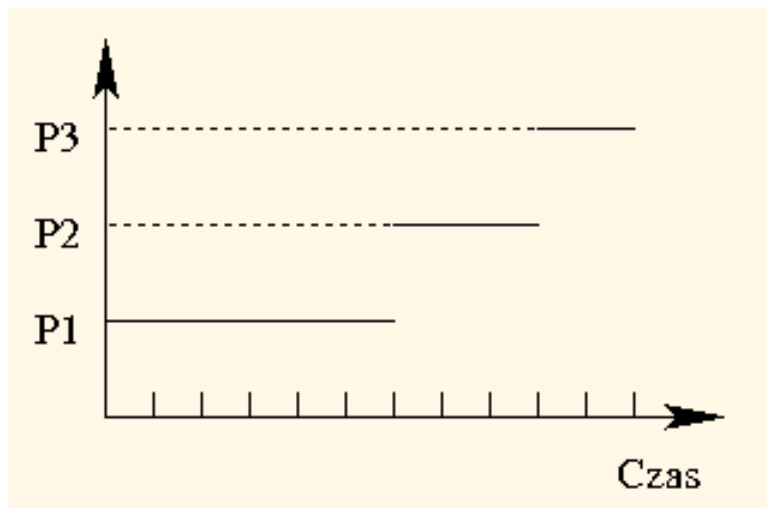
1. Co to jest przełączanie kontekstu?
2. Jaki jest związek pomiędzy zadaniami planisty i dyspozytora?
3. Jaką rolę pełni wywłaszczanie procesów?
4. Na czym polega sprawiedliwość w planowaniu procesów?
5. Jakie kryteria może optymalizować algorytm planowania?

Strategia FCFS

Najprostszym algorytmem planowania jest FCFS (*First-Come, First-Served*), czyli „pierwszy przyszedł pierwszy obsłużony.” FCFS jest strategią bez wywłaszczania. Procesy są wykonywane od początku do końca w takiej kolejności, w jakiej się pojawiły.

Proces	P1	P2	P3
Czas wykon.	6	3	2

Proces	P2	P3	P1
Czas wykon.	3	2	6

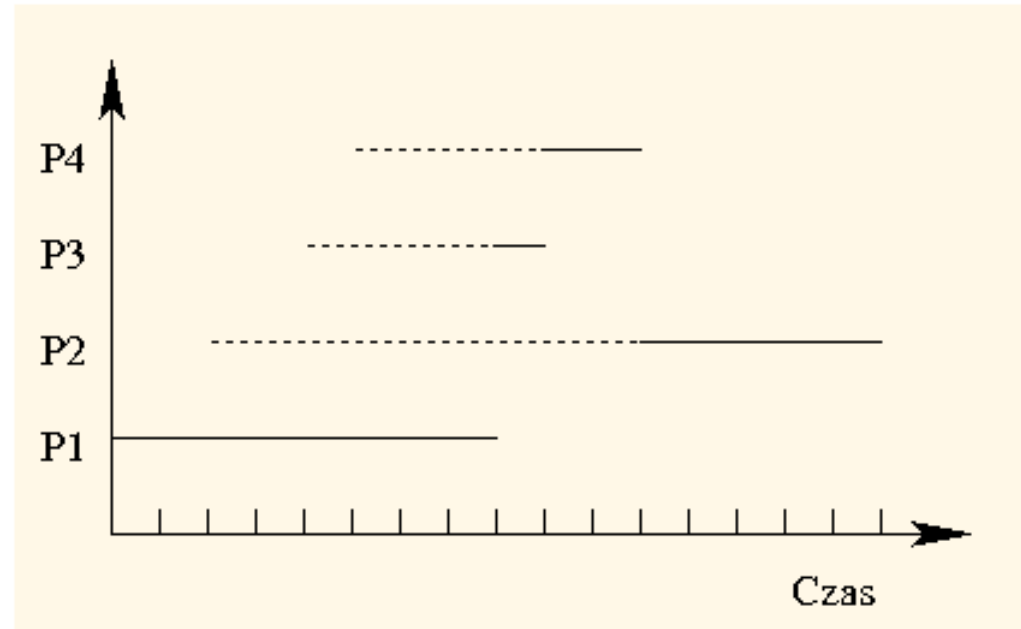


Warto zwrócić uwagę na to, że zadanie planisty przy tej strategii jest zredukowane do minimum (utrzymywanie kolejki). Zatem algorytm minimalizuje narzuty planowania. Planista nie musi też znać wymagań czasowych zadania.

Strategia SJF

Inną strategią planowania bez wywłaszczania jest SJF (*Shortest-Jobtime-First*), czyli „najpierw najkrótsze zadanie.” Spośród wszystkich gotowych uruchomiony zostaje proces, który ma najkrótszy całkowity czas wykonywania. Jego celem jest minimalizacja średniego czasu przetwarzania zadania. W tym celu jednak ten czas dla wszystkich procesów musi być znany planiście.

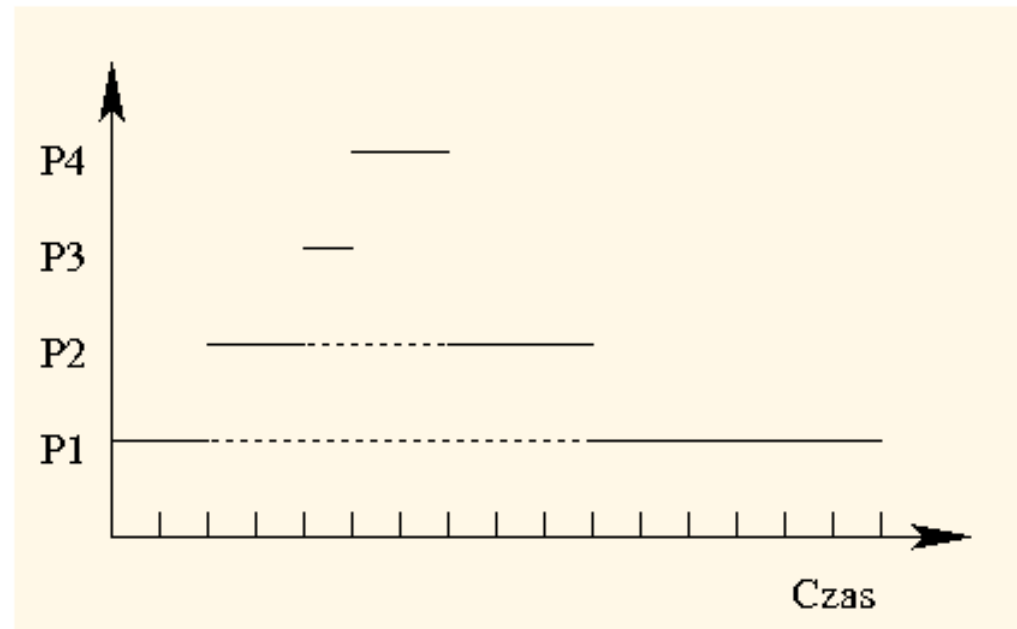
Proces	Czas uwoln.	Czas wykon.
P4	5	2
P3	4	1
P2	2	5
P1	0	8



Strategia SRTF

Odmianą algorytmu SJF z dodanym wywłaszczaniem jest SRTF (*Shortest-Remaining-Time-First*), czyli „najpierw proces o najkrótszym pozostałym czasie wykonania.” Zauważmy, że tu planista musi być wywołany, poza momentem zakończenia każdego procesu jak zwykle, jedynie w chwili pojawienia się nowego procesu, wybierając ten z najkrótszym pozostałym przewidywanym czasem obliczeń.

Proces	Czas uwoln.	Czas wykon.
P4	5	2
P3	4	1
P2	2	5
P1	0	8



Proces	P1	P2	P3	P4	P2	P1
Czas rozpoczęcia	0	2	4	5	7	10
Czas wykonany	2	2	1	2	3	6
Czas pozostały	6	3	0	0	0	0

Szacowanie fazy procesora

Zastanówmy się, jak oszacować czas następnej fazy procesora, potrzebny do realizacji strategii SJF i SRTF?

Planista z reguły nie wie ile będzie trwała kolejna faza procesora. Jednak gdy traktujemy czas życia procesu jako składający się z szeregu faz obliczeniowych, przedzielonych fazami I/O, to system zna długości jego poprzednich faz obliczeniowych. Zwykle fazy kolejne mają podobną charakterystykę jak poprzednie.

Jedną ze stosowanych metod szacowania następnej fazy procesora na podstawie długości poprzednich jest uśrednianie wykładnicze. Długość kolejnej fazy procesora szacujemy jako średnią ważoną długości poprzednich faz, przy czym wagi tworzą rosnący ciąg geometryczny (największą wagę ma ostatnia faza).

Zadania wsadowe i interakcyjne

W systemach operacyjnych mogą pojawiać się zadania o różnej charakterystyce. Dwie ważne grupy zadań stanowią: (i) **zadania obliczeniowe** (zwane również wsadowymi), które typowo wykonują obliczenia na procesorze i odwołują się głównie do pamięci RAM, i (ii) **zadania interakcyjne**, które typowo oczekują na dane z interfejsów I/O i wykonują krótkie obliczenia, lub kolejne transfery I/O, w odpowiedzi na otrzymane dane.

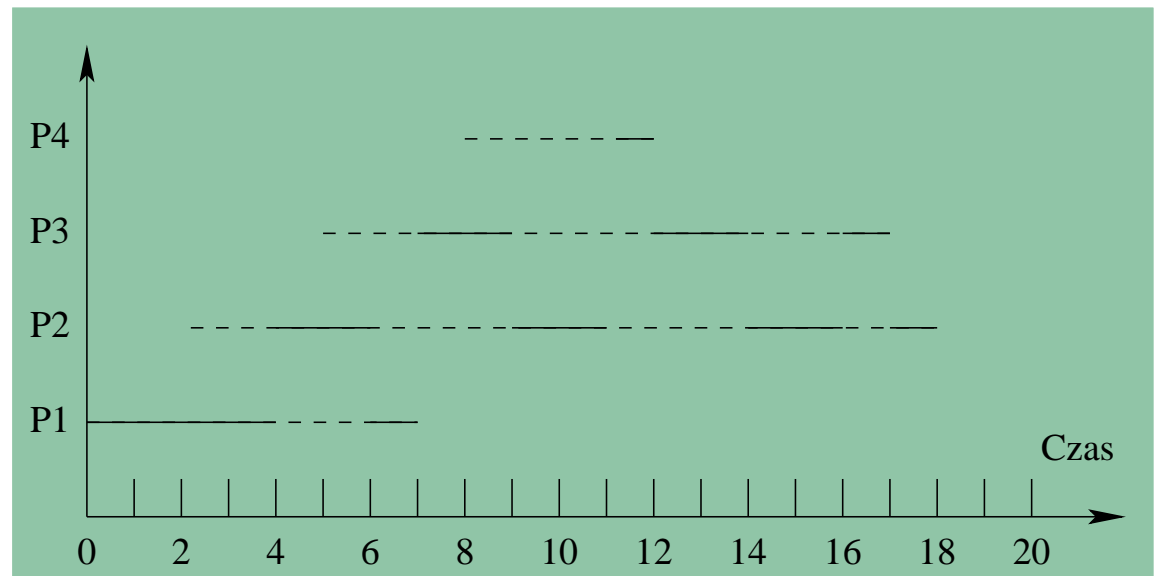
Zadania interakcyjne mogą być związane z pracą człowieka na komputerze typu PC, ale mogą to być również serwery odpowiadające na żądania napływające z sieci, albo przetwarzające dane z pomiarów, itp. Odmienny profil tych zadań powoduje, że przy ich szeregowaniu znaczenie mają różne kryteria, i lepiej nadają się do nich różne algorytmy szeregowania.

Omówione dotychczas algorytmy szeregowania FCFS, SJF, i SRTF optymalizują głównie wykorzystanie procesora i czas przetwarzania. Powoduje to, że dobrze nadają się do planowania zadań obliczeniowych, natomiast nie bardzo do zadań interakcyjnych, gdzie podstawowym kryterium powinny być: czas oczekiwania i przepustowość.

Planowanie rotacyjne

RR (*Round-Robin*), czyli planowanie rotacyjne, to strategia, w której każdy proces po kolei otrzymuje kwant czasu procesora do wykorzystania, po czym jest wywłaszczany. Zwykle jest to około 10-100 milisekund. Żaden proces nie czeka więc dłużej na procesor niż długość kwantu razy liczba procesów.

Proces	Czas uwoln.	Czas wykon.
P4	8	1
P3	5	5
P2	2	7
P1	0	5



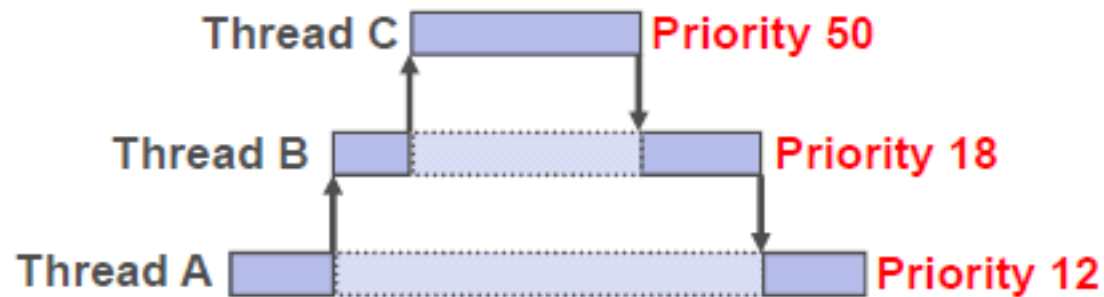
Proces	P1	P1	P2	P1	P3	P2	P4	P3	P2	P3	P2
Czas rozpoczęcia	0	2	4	6	7	9	11	12	14	16	17
Czas wykonany	2	2	2	1	2	2	1	2	2	1	1
Czas pozostały	3	1	5	0	3	3	0	1	1	0	0

Planowanie rotacyjne jest zupełnie inną strategią planowania od przedstawionych wcześniej. **Nadaje się ono głównie do planowania zadań interakcyjnych, gdzie najważniejszym kryterium jest krótki czas reakcji.**

Planowanie rotacyjne powoduje znacznie więcej przełączeń kontekstu niż poprzednie algorytmy planowania, jest więc dość kosztowną strategią. Kwant czasu przydzielany procesom musi być co najmniej o rząd wielkości większy niż czas przełączenia kontekstu. W przeciwnym wypadku narzuty generowane przez planowanie będą znaczące, a nawet mogą przewyższyć rzeczywistą pracę wykonywaną przez system na obliczenia procesów.

Planowanie priorytetowe

Jeżeli dla każdego procesu określimy liczbę reprezentującą jego **priorytet**, to możliwe jest **planowanie priorytetowe**. Procesor jest przydzielany procesowi, który ma największy priorytet. Możemy rozróżnić wersję wywłaszczeniową, gdy proces jest usuwany z procesora natychmiast po pojawieniu się procesu o wyższym priorytecie, oraz wersję niewywłaszczeniową, gdy procesy okresowo dobrowolnie zwalniają procesor, i żaden nie jest wywłaszczany.



Planowanie priorytetowe jest atrakcyjne ponieważ pozwala łatwo określić, który proces będzie wykonywany przed którym. Dlatego bywa chętnie stosowane w systemach gdzie konieczne jest zagwarantowanie określonej kolejności wykonania procesów, jak systemy czasu rzeczywistego, systemy wbudowane, systemy wojskowe, itp.

Jednak prawidłowe zdefiniowanie priorytetów dla określonych wymagań jest proste tylko gdy system jest mały, typu kilku do kilkunastu procesów. Gdy liczba procesów jest większa, określenie właściwych wartości priorytetów zaczyna być problemem.

Priorytety statyczne i dynamiczne

Priorytety mogą być przydzielone zadaniom sztywno (statycznie), według schematu określonego przez programistę lub administratora systemu. Jednak to da się zrobić jedynie w odniesieniu do systemów zamkniętych, w których istnieje określona pula zadań o dobrze znanych własnościach.¹

W systemach otwartych priorytety muszą być przydzielane pojawiającym się nieznanym zadaniom. Inteligentny planista może przydzielać zadaniom wstępne priorytety według jakiegoś przyjętego schematu, a następnie dynamicznie aktualizować je na podstawie obserwacji ich pracy.

Sensowną strategią byłoby rozpoznawanie zadania jako obliczeniowego, i obniżanie jego priorytetu, lub jako interakcyjnego, i podwyższanie priorytetu. Wtedy maksimum zasobów obliczeniowych będzie przydzielane procesom interakcyjnym, natomiast procesy obliczeniowe będą wykonywane tylko w przerwach, gdy wszystkie procesy interakcyjne dobrowolnie zwolnią procesor. W ten sposób, planowanie priorytetowe nadaje się zarówno do planowania procesów obliczeniowych, interakcyjnych, jak i mieszanych.

¹A nawet w takich systemach statyczny przydział priorytetów jest nietrywialnym zadaniem, gdy pojawia się duża liczba procesów o złożonych zadaniach.

Strategie złożone

W systemie planującym wykonanie zadań algorytmem priorytetowym może pojawić się wiele zadań o tym samym priorytecie. Wcale nie musi to być przypadek ani sytuacja wyjątkowa. Zadania zwykle mają priorytety przydzielane według pewnego schematu. Jeśli więc zostanie uruchomiona pewna liczba podobnych zadań to będą one mieć równe priorytety.

A wtedy, spośród grupy zadań, które wszystkie mają najwyższy priorytet, które powinno zostać uruchomione pierwsze? Przypadkowy ich wybór zwykle nie jest dobrym rozwiązaniem, bo jego wynikiem będzie nierównomierny czas oczekiwania na wykonanie zadań, które teoretycznie powinny być traktowane jednakowo.

Dobrym rozwiązaniem jest zastosowanie drugorzędnego algorytmu planowania, który będzie wybierał zadania spośród tych o najwyższym priorytecie. Ponieważ powinny one być równoważne, i równo traktowane, dlatego w roli drugorzędnego algorytmu planowania często stosowany jest RR.

Wielopoziomowe kolejki

Przedstawiliśmy strategie dobre do szeregowania procesów obliczeniowych (SJF i SRTF) oraz procesów interakcyjnych (RR). Jeśli w systemie będą obecne procesy obu rodzajów, to system stosujący wyłącznie jedną wybraną strategię będzie przetwarzał część procesów nieoptymalnie.

Zamiast tego, **można użyć dla każdej z tych grup procesów innej, odpowiedniej dla danej grupy, strategii planowania**. Procesy obliczeniowe mogą być obsługiwane strategią SJF, ponieważ w przypadku pracy wsadowej ważna jest minimalizacja średniego czasu oczekiwania. Natomiast procesy interakcyjne mogą być obsługiwane strategią RR, ponieważ w przypadku pracy interakcyjnej istotna jest minimalizacja czasu reakcji.

Typ procesu może być określony przez deklarację użytkownika, albo przez rozpoznanie jego charakterystyki przez system, jak w przypadku dynamicznych priorytetów.

Należy jeszcze określić wzajemny stosunek obu tych grup procesów. Oczywiście procesy interakcyjne powinny być uprzywilejowane, stąd często nazywa się je procesami **pierwszoplanowymi**, albo procesami **czoła**, a procesy obliczeniowe procesami **tła**. W najprostszym przypadku można przyjąć, że procesy czoła będą wykonywane zawsze gdy są gotowe, w oczekiwaniu, że pozostawią one część cykli obliczeniowych niewykorzystanych, umożliwiając wykonywanie zadań tła.

Partycjonowanie

Opisane wielopoziomowe kolejki z uprzywilejowanymi procesami czoła mogą w pewnych przypadkach doprowadzić do zagłodzenia procesów tła. Takie ryzyko może być akceptowalne w pewnych przypadkach, lecz ogólnie zjawisko jest niepożądane w systemach operacyjnych.

Można temu zapobiec, przyjmując stały podział czasu między kolejki procesów tła i czoła, np. 90% dla czoła i 10% dla tła. Jest to strategia zwana **partycjonowaniem**.

Partycjonowanie może samo w sobie być uważane za algorytm planowania. Jednak podobnie jak planowanie priorytetowe, może ono wydzielić grupę zadań do wykonania bez możliwości wyboru zadania w ramach grupy. Jest to zatem algorytm wysokiego poziomu, który należy uzupełnić o dodatkową strategię, pozwalającą szeregować zadania w ramach poszczególnych partycji.

Wadą planowania przez partycjonowanie jest potencjał do marnowania wolnej mocy procesora. Jeśli zadania z jednej partycji nie wykorzystują wszystkich cykli procesora dostępnych w tej partycji, to zadania z innej partycji nie mogą ich wykorzystać. Jest możliwe **partycjonowanie adaptacyjne** udostępniające wolne cykle procesora z jednej partycji zadaniom z innej partycji.

Krótkie podsumowanie — pytania sprawdzające

Odpowiedz na poniższe pytania:

1. Porównaj własności strategii planowania FCFS i SJF (przewagi jednej i drugiej).
2. Czym różni się zachowanie zadań wsadowych od interakcyjnych?
3. Jakie kryteria optymalizuje algorytm planowania rotacyjnego (RR) i kosztem jakich parametrów się to odbywa?
4. Dlaczego planowanie priorytetowe typowo łączy się z inną strategią planowania?
5. W jakim celu stosuje się planowanie procesów z wielopoziomowymi kolejkami?
6. Jakie są zalety i wady zastosowania partycjonowania w planowaniu?

Inne przesłanki w zagadnieniach planowania

Warto zwrócić uwagę na fakt, że mogą wystąpić okoliczności, które nie są znane planiście, ale mogą zasadniczo wpłynąć na efektywność pracy systemu.

Jeśli istnieje grupa procesów zorientowanych na obliczenia (*compute-bound*), i grupa procesów zorientowanych na operacje wejścia-wyjścia (*I/O-bound*), i planista zaplanuje najpierw wykonywanie tych pierwszych, to będą one rywalizowały o dostęp do procesora, a w tym czasie dysk, i inne urządzenia I/O, będą bezczynne. Gdy następnie uruchomiona zostanie druga grupa procesów, to będą one stale blokowały się na operacjach I/O i zwalniały procesor, który będzie mało obciążony, podczas gdy podsystem dyskowy będzie pracował na 100%.

Przewidując to, planista (długoterminowy) mógłby uruchomić najpierw tylko część procesów pierwszej i drugiej grupy pozwalając aby procesy obliczeniowe wykonywały się w chwilach zwolnienia procesora przez procesy I/O.

Jednak rzadko kiedy planista ma dostęp do takich informacji. Natomiast w systemach, gdzie wykonują się dobrze zaprojektowane aplikacje, programista i/lub administrator mogą mieć możliwość ingerencji w algorytm planowania, w celu poprawy wydajności pracy systemu (patrz mechanizm i funkcja systemowa *nice*).

Optymalizacja w planowaniu

Algorytmy planowania optymalizują sposób wykonywania procesów przez system:

- z punktu widzenia systemu, co oznacza minimalizację strat, czyli przełączeń kontekstu; skrajnym przypadkiem są strategie FCFS i SJF, które nie wykonują żadnych przełączeń kontekstu niepotrzebnych z punktu widzenia systemu; wszystkie strategie z wywłaszczaniem idą na kompromis, akceptując pewną liczbę „niepotrzebnych” przełączeń kontekstu,
- z punktu widzenia użytkownika, co oznacza wybór właściwych procesów do wykonywania, gdzie przez „właściwe” rozumie się:
 - procesy, które wiadomo, że powinny być wykonywane pierwsze, zgodnie z ideą danego algorytmu, np. procesy o najwyższym priorytecie w planowaniu priorytetowym, albo procesy interakcyjne w planowaniu z wielopoziomowymi kolejkami,
 - równomierne wykonywanie wszystkich „pozostałych” procesów, tzn. takich, co do których system nie ma żadnych preferencji wykonywania.

„Sprawiedliwość” w planowaniu

Rozważmy jak zasadę „sprawiedliwości”, czyli egalitaryzmu, stosując poznane algorytmy planowania.

- Algorytm FCFS realizuje pewien specyficzny rodzaj sprawiedliwości, mianowicie prawo pierwszeństwa. Nie jest ono jednak zgodne z zasadą egalitaryzmu i nie uznajemy go za sprawiedliwość.
- Algorytmy SJF i SRTF faworyzują zadania krótkie, mogą doprowadzić do zagłodzenia zadań długich, i nie są sprawiedliwe.
- Algorytm RR jest czystą realizacją zasady sprawiedliwości, kosztem maksymalnych strat na przełączanie kontekstu.
- Strategię opartą na priorytetach można by uznać za częściowo sprawiedliwą, jeśli procesy o równych priorytetach byłyby szeregowane algorytmem sprawiedliwym, takim jak RR. Byłaby to jednak sprawiedliwość „kastowa” dzieląca obywateli na „równych” i „równiejszych.”
Szeregowanie priorytetowe uznajemy więc za niesprawiedliwe.

Podsumowanie i inne rodzaje planowania

Istnieje grupa strategii planowania szczególnie nadających się do (optymalizujących) wykonywania zadań obliczeniowych: FCFS, SJF, SRTF.

Inna grupa strategii jest bardziej odpowiednia do zadań interakcyjnych: RR i strategie mieszane (np. w połączeniu z priorytetami, albo wielopoziomowe).

Omawiane tu strategie powstały w odniesieniu do systemów jednoprocessorowych (na co zresztą wskazuje polskie określenie: szeregowanie). Mogą być one również stosowane do systemów wieloprocessorowych i wielordzeniowych, jakkolwiek istnieją nieomówione tu wyspecjalizowane wersje algorytmów planowania dla takich systemów, biorące pod uwagę ich specyfikę.

Odmienne wymagania prezentują również systemy czasu rzeczywistego. Specjalizowane systemy operacyjne czasu rzeczywistego (*RTOS — Real Time Operating System*) mają ogólnie swoją specyfikę, odróżniającą je od systemów GPOS. Na przykład, obce są im obowiązujące w GPOS zasady sprawiedliwości. W systemie RTOS dopuszczalne jest, by jeden proces wysokiego priorytetu zdominował cały system, włącznie z wyłączeniem jądra systemu.

Krótkie podsumowanie — pytania sprawdzające

Odpowiedz na poniższe pytania:

1. Które strategie planowania nadają się do planowania procesów obliczeniowych, które do interakcyjnych, a które można stosować dla mieszanej grupy procesów?
2. Które strategie planowania mają charakter niewywłaszczeniowy, które są typowo wywłaszczeniowe, a które można stosować zarówno z wywłaszczeniem jak i bez niego?

Ciekawe materiały

Wykład wideo z UC Berkeley o planowaniu procesów – prawie 1.5 godziny

http://www.youtube.com/watch?v=a0DdUp_eHQs