Systemy operacyjne

wykład 1 - System operacyjny i jego zadania

dr Marcin Ziółkowski

Instytut Matematyki i Informatyki
Akademia im. Jana Długosza w Częstochowie

3 marca 2016 r.

PLAN WYKŁADU

1. Historia i zadania systemów operacyjnych
2. Interpreter poleceń i jego wykorzystanie
3. Struktura systemu komputerowego oraz systemu operacyjnego
4. Procesy, zarządzanie procesami, synchronizacja i zakleszczenia
5. Zarządzanie pamięcią
6. System plików
7. Zarządzanie dyskiem
8. Przykłady systemów operacyjnych

LITERATURA

[1] A.Silberschatz, P.B.Galvin, Podstawy systemów operacyjnych, WNT 2000.
[2] M.J.Bach, Budowa systemu operacyjnego Unix, WNT 1995.
[3] M.J.Rochkind, Programowanie w systemie Unix dla zaawansowanych, WNT 1997.

Zadania systemu operacyjnego

System operacyjny jest istotną częścią każdego systemu komputerowego. Istnieje wiele różnych definicji systemu operacyjnego. Jednak co do zasady można się zgodzić, że system operacyjny powinien spełniać następujące zadania:

- pośredniczyć pomiędzy użytkownikiem a sprzętem komputerowym
- udostępniać programom maszynę wirtualną
- przydzielać sprzętowe zasoby programom komputerowym
- nadzorować pracę programów
- tworzyć bezpieczne i wygodne środowisko pracy dla użytkownika
- przechowywać rozmaite informacje dla użytkownika

Cele systemu operacyjnego

Z opisanych zadań wynikają następujące cele systemu operacyjnego:

1. umożliwienie użytkownikom uruchamiania programów i ułatwienie wykonywanych przez nich zadań
2. stworzenie użytkownikom wygodnego i funkcjonalnego środowiska pracy
3. efektywne wykorzystanie zasobów systemu komputerowego

System operacyjny - definicja

Najczęściej przyjmuje się zatem następującą definicję:

SYSTEM OPERACYJNY - DEFINICJA

SYSTEM OPERACYJNY - całość oprogramowania realizującego cele związane z umożliwieniem użytkownikom uruchamiania programów i ułatwieniem wykonywanych przez nich zadań, stworzeniem użytkownikom wygodnego i funkcjonalnego środowiska pracy oraz efektywnym wykorzystaniem zasobów systemu komputerowego. Oprogramowanie takie jest dostarczane zwykle razem ze sprzętem przez dostawcę sprzętu komputerowego.

Historia systemów operacyjnych - systemy wsadowe

Pierwsze komputery były dużymi i drogimi urządzeniami niedostępnymi dla zwykłego użytkownika. Wyposażone były w czytniki kart, przewijaki taśm, konsole oraz drukarki. Ze względu na ich cenę oraz koszty eksploatacji priorytetem było efektywne wykorzystanie czasu pracy komputera. Pracę komputera nadzorował wyspecjalizowany operator. Zwykli użytkownicy przekazywali swoje zadania - programy i dane wejściowe (w postaci kart perforowanych), aby następnie odebrać wyniki (w postaci wydruków). Operator łączył zadania o podobnych wymaganiach w tzw. wsady. Systemy operacyjne wówczas były bardzo proste, ich rola sprowadzała się do wczytywania kolejnych zadań i wykonywania ich. Miarą wydajności systemu wsadowego był czas obrotu zadania czyli czas między złożeniem zadania, a otrzymaniem wyników.

Spooling

We wczesnych systemach wsadowych wykorzystanie procesora było bardzo nieefektywne. Z uwagi na fakt, że prędkość procesora była o kilka rzędów wielkości większa niż prędkości urządzeń wejścia-wyjścia, procesor często pozostawał bezczynny, ponieważ operacje wejścia-wyjścia w nieznacznym stopniu angażowały procesor. W momencie pojawienia się dysków magnetycznych - szybkich nośników informacji pojawiła się nowa technologia, która pozwalała na zwiększenie wydajności systemu. Spooling polegał na tym, że jednocześnie odbywały się trzy rzeczy:

- zapisywanie na dysku nowych zadań, które zostały wczytane przez czytnik
- pobieranie przez procesor z dysku kolejnych zadań i wykonywanie ich oraz zapisywanie wyników na dysku
- drukowanie wyników zakończonych zadań na drukarkach

Nowym zadaniami systemu operacyjnego stały się natomiast synchronizacja wczytywania zadań, wypisywania wyników oraz zarządzanie informacją zapisaną na dysku.

Mechanizm spoolingu zwiększał wydajność procesora, jednak nadal nie był wykorzystywany efektywnie - na przykład w momentach wykonywania operacji dyskowych. Dalszą poprawę uzyskano przez wprowadzenie wieloprogramowości, która polega na tym, że w pamięci operacyjnej komputera znajduje się wiele działających programów. Program uruchomiony i załadowany do pamięci operacyjnej komputera nazywa się procesem. Gdy jeden z procesów czeka na zakończenie operacji wejścia-wyjścia, procesor nie musi być bezczynny, ponieważ może wówczas obsługiwać inny z procesów.

Jednak, ponieważ nie wszystkie zadania, które zostały wczytane, mogą zmieścić się w pamięci operacyjnej, przed systemami pojawiają się kolejne zadania:

- zarządzanie pamięcią (oraz pilnowanie, aby nie zachodziły konflikty w przydziale pamięci dla różnych procesów)
- przydzielanie procesom urządzeń zewnętrznych
- szeregowanie krótkoterminowe (jeżeli procesor jest bezczynny, należy wybrać jeden z procesów, który znajduje się w pamięci i jest gotowy do wykonania)
- szeregowanie długoterminowe (jeżeli pamięć operacyjna jest wolna, należy wczytać z dysku kolejne zadanie oczekujące na wykonanie)

Systemy z podziałem czasu

Wieloprogramowość poprawiła efektywność wykorzystania komputera, jednak nadal czas pracy programistów był bardzo nieefektywnie wykorzystywany. Użytkownicy nadal nie mieli bezpośredniej styczności z komputerem, a jedynie wsadowo przetwarzali zadania. Sytuacja zmieniła się, gdy czytniki kart zastąpiono terminalami, przy których mogli pracować użytkownicy. Praca stała się interaktywna. Podział czasu polegał na tym, że procesor był przydzielany procesom w małych porcjach czasowych tzw. kwantach. Po zakończeniu się kwantu czasowego procesor przełączał się na wykonywanie kolejnego procesu. Kwanty czasu były na tyle małe, że użytkownik ma wrażenie płynnej pracy, poza tym - jeden użytkownik nie obciąża przez cały czas procesora, powstaje więc wrażenie, że każdy z użytkowników ma dostęp do dużej części mocy obliczeniowej komputera.

Systemy z podziałem czasu

Przy interaktywnej komunikacji charakterystyką, która lepiej oddaje wydajność systemu jest czas reakcji - czas, jaki upływa od wykonania przez użytkownika jakiejś akcji (naciśnięcia klawisza, wprowadzenia polecenia, kliknięcia myszą itp.) do zareagowania przez proces na tę akcję. Oczywiście czas reakcji obejmuje czas oczekiwania na przydzielenie procesowi do pracy procesora.
Z uwagi na interaktywny udział użytkowników, w systemach z podziałem czasu pojawiły się nowe potrzeby i wymagania. Jedną z najważniejszych było stworzenie mechanizmu przechowywania danych i programów (systemu plików). Bardzo często komputery z przetwarzaniem wsadowym miały systemy operacyjne wzbogacane o opcję podziału czasu.

Komputery osobiste

Wraz z pojawieniem się tańszych komputerów osobistych przeznaczonych głównie do pracy dla jednego użytkownika zmieniła się trochę hierarchia zadań wykonywanych przez systemy operacyjne. Wydajność komputera nie była najważniejsza dla szeregowego użytkownika. Najistotniejsza stała się wygoda pracy użytkownika. Ważne z punktu użytkownika stało się więc na przykład, jakie narzędzia dodatkowe znajdują się w systemie, jaki ma interfejs graficzny itd.
W komputerach osobistych początkowo nie było wieloprogramowości, jednak z czasem stały się również bardzo wydajnymi maszynami. W ogóle można zauważyć ciekawe zjawisko - w miarę postępu technologii, rozwiązania wydajnościowe znane z komputerów "cięższej kategorii" są stosowane w "lekkich" komputerach.

Rodzaje systemów operacyjnych

Wśród różnych systemów operacyjnych wyróżniamy między innymi:

- systemy równoległe
- systemy rozproszone
- systemy czasu rzeczywistego

Systemy równoległe

Większość systemów komputerowych to systemy wyposażone w jeden procesor. Systemy równoległe to systemy wyposażone w wiele procesorów, które mogą równolegle wykonywać obliczenia. Procesory te mogą być ze sobą **ściśle powiązan**e (gdy współdzielą np. magistralę, pamięć, urządzenia zewnętrzne) lub **luźno powiązane** (gdy każdy z nich posiada własną magistralę i pamięć, tworzy pewien podsystem komputerowy, który może komunikować się z pozostałymi podsystemami poprzez szybkie linie komunikacyjne). Spośród systemów ze ścisłym powiązaniem procesorów wyróżniamy **systemy z symetrycznymi procesorami** i **asymetrycznymi procesorami**. Pierwszy z systemów to system z układem równoprawnych procesorów centralnych, które współdzielą zegar, magistralę i pamięć i zawierają pewne narzędzie ich synchronizacji. Takie systemy są najpopularniejszymi systemami równoległymi ogólnego przeznaczenia.

Systemy równoległe

Drugi z rodzajów to system, w którym jest jeden procesor centralny oraz kilka wyspecjalizowanych procesorów pomocniczych, które wykonują zadania rozdzielone przez procesor centralny. Takie systemy są stosowane w wyspecjalizowanych serwerach. Systemy wieloprocesorowe znacznie zwiększają moc obliczeniową komputerów i są korzystne również ze względów ekonomicznych (cena szybszych procesorów, awarie jednego procesora nie doprowadzają do bezczynności całego systemu).

Systemy rozproszone

Systemy rozproszone są szczególnym przypadkiem systemów równoległych z luźno powiązanymi procesorami. Są to systemy, w których wiele komputerów jest połączonych w jedną sieć i tworzy jeden system. Użytkownik dowolnego komputera widzi wtedy system rozproszony jako jedną wspólną całość. Zaletami takich systemów są:

- niezawodność - przy awarii jednego komputera, pozostałe mogą częściowo przejąć jego zadania
- współdzielenie zasobów - przechowywanie danych użytkowników na wspólnych dyskach (koszty)
- zwiększenie mocy obliczeniowej - równoważenie obciążeń komputerów
- nowe usługi (możliwe do zrealizowania tylko przez sieć komputerów - poczta elektroniczna)

Systemy rozproszone

Rozproszenie wszystkich funkcji systemu operacyjnego jest trudne do realizacji, więc często tylko niektóre funkcje są rozproszone - na przykład rozproszony system plików. Najczęściej spotykana architektura systemów rozproszonych to architektura client-server - wiele stacji klienckich korzysta z usług jednego serwera. W pewnych specjalnych przypadkach, gdy ważniejsza jest od wydajności niezawodność systemu - tworzy się grupy komputerów dublujących swoje zadania, widocznych jako jeden system komputerowy - clusters. W przypadku awarii jednego z nich pozostałe z komputerów przejmują jego zadania.

Systemy czasu rzeczywistego

Systemy czasu rzeczywistego to systemy, które działają w warunkach pewnych ograniczeń czasowych np. reagowania na pewne zdarzenia z ograniczonym opóźnieniem czasowym. Dzielimy je na **systemy z bezwzględnymi ograniczeniami czasowymi** oraz **systemy z łagodnymi ograniczeniami czasowymi**. Pierwsza z klas to systemy sterujące zwykle jakimiś urządzeniami technicznymi lub procesami technologicznymi. Wymaga się od nich reagowania natychmiastowego na niektóre zdarzenia mogące doprowadzić do sytuacji niebezpiecznych dla urządzeń oraz obsługujących je ludzi. W przypadku takich systemów spełnienie tych wymogów (wymagań bezpieczeństwa) jest ważniejsze od ich wydajności. Druga z klas to systemy, dla których niespełnienie wymogów prowadzi do pogorszenia jakości usług - tu przykładem są urządzenia multimedialne.