

12.3 АЛГОРИТМИ ПЛАНУВАННЯ, ЗАСНОВАНІ НА ПРІОРИТЕТАХ

Іншою важливою концепцією, що лежить в основі багатьох витісняючих алгоритмів планування, є пріоритетне обслуговування. Пріоритетне обслуговування передбачає наявність у потоків деякої відомої характеристики – пріоритету, на підставі якої визначається порядок їх виконання.

Пріоритет – це число, що характеризує ступінь привілейованості потоку при використанні ресурсів обчислювальної машини, зокрема процесорного часу: чим вищий пріоритет, тим вищі привілеї, тим менше часу буде проводити потік в чергах.

Пріоритет може виражатися цілим або дробовим, додатнім або від'ємним числом. У деяких ОС прийнято, що пріоритет потоку тим вищий, чим більше число (в арифметичному сенсі), що позначає пріоритет. В інших системах, навпаки, чим менше число, тим вище пріоритет.

Так, в UNIX і багатьох інших системах великі значення пріоритетів відповідають процесам з низьким пріоритетом. Деякі системи, такі як OS/2 або Windows, використовують зворотне значення: більше значення вказує на вищий пріоритет.

У більшості операційних систем, що підтримують потоки, пріоритет потоку безпосередньо пов'язаний з пріоритетом процесу, в рамках якого виконується даний потік. Пріоритет процесу назначається операційною системою при його створенні. Значення пріоритету включається в описувач процесу і використовується при призначенні пріоритету потокам цього процесу. При призначенні пріоритету новоствореному процесу ОС враховує, чи є цей процес системним чи прикладним, який статус користувача, що запустив процес, чи була явна вказівка користувача на присвоєння процесу певного рівня пріоритету. Потік може бути ініційований не тільки командою користувача, але і в результаті виконання системного виклику іншим потоком. У цьому випадку при призначенні пріоритету новому потоку ОС повинна брати до уваги значення параметрів системного виклику.

У багатьох ОС передбачається можливість зміни пріоритетів протягом життя потоку. Зміни пріоритетів можуть відбуватися з ініціативи самого потоку, коли він звертається з відповідним викликом до операційної системи. Пріоритет може

призначатися директивно адміністратором системи в залежності від важливості, або обчислюватися самою ОС за певними правилами. Пріоритет може також змінюватися з ініціативи користувача, коли він виконує відповідну команду. Крім того, ОС сама може змінювати пріоритети потоків в залежності від ситуації, що складається в системі. В останньому випадку пріоритети називаються динамічними.

Від того, які пріоритети призначені потокам, істотно залежить ефективність роботи всієї обчислювальної системи. У сучасних ОС, щоб уникнути розбалансування системи, яке може виникнути при неправильному призначенні пріоритетів, можливості користувачів впливати на пріоритети процесів і потоків намагаються обмежувати. Права підвищувати пріоритети потокам дозволено робити тільки адміністраторам.

Як приклад розглянемо схему призначення пріоритетів потокам, прийняту в операційній системі Windows NT (рис. 12.6). В системі визначено 32 рівні пріоритетів і два класи потоків – потоки реального часу і потоки зі змінними пріоритетами. Діапазон від 1 до 15 включно відведений для потоків зі змінними пріоритетами, а від 16 до 31 – для критичніших потоків реального часу (пріоритет 0 зарезервований для системних цілей).

При створенні процесу він у залежності від класу отримує за замовчуванням базовий пріоритет у верхній або нижній частині діапазону. Базовий пріоритет процесу в подальшому може бути підвищений або знижений операційною системою. Спочатку потік отримує значення базового пріоритету з діапазону базового пріоритету процесу, в якому він був створений.



Рисунок 12.6 – Схема призначення пріоритетів у Windows NT

Нехай, наприклад, значення базового пріоритету деякого процесу дорівнює K . Тоді усі потоки даного процесу отримають базові пріоритети з діапазону $[K-2, K+2]$. Звідси видно, що, змінюючи базовий пріоритет процесу, ОС може впливати на базові пріоритети його потоків.

У Windows NT з плином часу пріоритет потоку може відхилятися від базового пріоритету потоку, причому ці зміни можуть бути не пов'язані зі змінами базового пріоритету процесу. ОС може підвищувати пріоритет потоку (який у цьому випадку називається динамічним) в тих випадках, коли потік не повністю використав відведений йому квант, або знижувати пріоритет, якщо квант був використаний повністю. ОС нарощує пріоритет в залежності від того, якого типу подія не дала потоку повністю використати квант.

Зокрема, ОС підвищує пріоритет у більшій мірі потокам, які очікують введення з клавіатури (інтерактивним додаткам) і в меншій мірі потокам, які виконують дискові операції. Саме на основі динамічних пріоритетів здійснюється планування потоків. Початковою точкою відліку для динамічного пріоритету є значення базового пріоритету потоку. Значення динамічного пріоритету потоку обмежено знизу його базовим пріоритетом, верхнім же кордоном є нижня межа діапазону пріоритетів реального часу.

Існують два різновиди пріоритетного планування: обслуговування з відносними пріоритетами і обслуговування з абсолютними пріоритетами.

В обох випадках вибір потоку на виконання з черги готових здійснюється однаково: вибирається потік, що має найвищий пріоритет. Однак зміни активного потоку вирішується по-різному. У системах з відносними пріоритетами активний потік виконується до тих пір, поки він сам не покине процесор, перейшовши в стан очікування (або ж станеться помилка, або потік завершиться). На рис. 12.7, а показаний граф станів потоку в системі з відносними пріоритетами.

У системах з абсолютними пріоритетами виконання активного потоку переривається крім зазначених вище причин, ще за однієї умови: якщо в черзі готових потоків з'явився потік, пріоритет якого вище пріоритету активного потоку. Перерваний потік переходить в стан готовності (рис. 12.7, б).

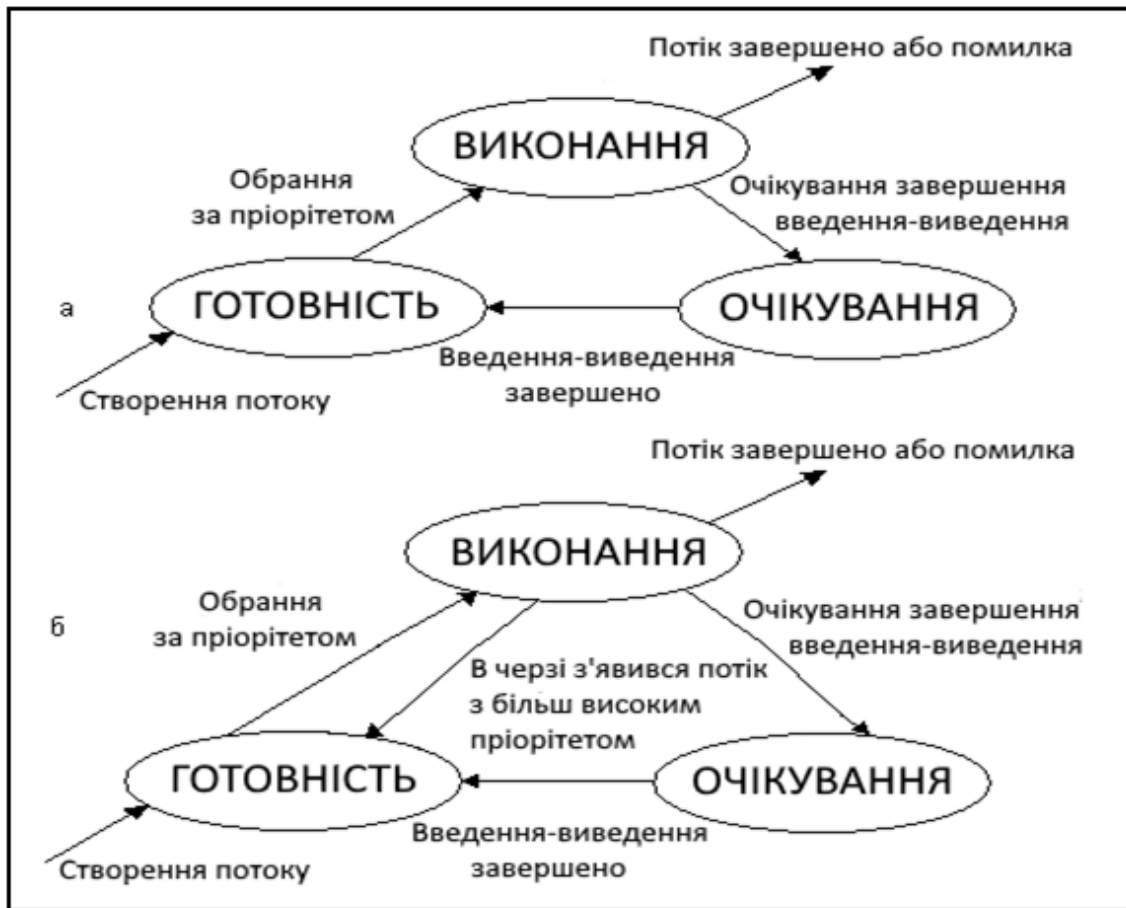


Рисунок 12.7 – Графи станів потоків в системах з відносними (а) і абсолютними пріоритетами (б)

У системах, в яких планування здійснюється на основі відносних пріоритетів, мінімізуються витрати на перемикання процесора з однієї роботи на іншу. З іншого боку, тут можуть виникати ситуації, коли одна задача займає процесор довгий час. Ясно, що для систем поділу часу і реального часу така дисципліна обслуговування не підходить, інтерактивний додаток може чекати своєї черги годинами, поки обчислювальної задачі не потрібно введення-виведення. А ось в системах пакетної обробки (в тому числі у відомої ОС OS/360) відносні пріоритети використовуються широко.

У системах з абсолютними пріоритетами час очікування потоку в чергах може бути зведений до мінімуму, якщо йому призначити найвищий пріоритет. Такий потік буде витіснити з процесора всі інші потоки (крім потоків, що мають такий же найвищий пріоритет). Це робить планування на основі абсолютних пріоритетів відповідним для систем управління об'єктами, в яких важлива швидка реакція на подію.