Transakcje rozproszone, awarie i kontrola współbieżności

Robert A. Kłopotek

r.klopotek@uksw.edu.pl

Wydział Matematyczno-Przyrodniczy. Szkoła Nauk Ścisłych, UKSW

Rozproszone odtwarzanie po awarii – wybrane problemy

- W niektórych przypadkach trudno jest nawet sprawdzić czy stanowisko uległo awarii, bez wymiany określonej liczby komunikatów pomiędzy różnymi stanowiskami
- Załóżmy, że stanowisko X wysyła komunikat do Y i spodziewa się odpowiedzi od Y ale jej nie otrzymuje. Istnieje kilka możliwości:
  - Wystąpił błąd komunikacji i komunikat nie dotarł do Y
  - Wystąpiła awaria stanowiska Y i nie może ono odpowiedzieć
  - Y wysłało odpowiedź, ale nie dotarła ona do X

Znajdowanie się w awarii

- W przypadku zagubienia danych, system powinien zadbać o retransmisje
- W przypadku awarii łącza, system jest w stanie znaleźć nową, optymalną drogę przesyłania danych
- Maszyna, która startuje ponownie powinna zadbać o to, aby zaktualizować dane, które zostały zmienione w trakcie trwania awarii, z reguły wstrzymuje się działanie systemy na czas naprawienia awarii

Problemy z transakcjami w systemach rozproszonych:

- Faza potwierdzenia (commit) może trwać długo. Niepowodzenie (awaria) w tej fazie może spowodować niespójność bazy danych i przetwarzania. Stąd nowe protokoły przetwarzania transakcji w systemie rozproszonym: 2PC, 3PC.
- Zakleszczenie: wykrywanie i usuwanie jest znacznie trudniejsze, gdyż wymaga przesyłania w sieci informacji która transakcja czeka na którą.
- Ziarnistość mechanizmu transakcji: bezpośrednio związana z dostępnością rozproszonej bazy danych (liczbą jednocześnie pracujących użytkowników).
- Wydajność (performance): złożone protokoły powodują jej zmniejszenie; stąd 2PC i 3PC są często nieadekwatne.
- Długie transakcje: konieczność zmniejszenia poziomu izolacji aby umożliwić dostęp do zablokowanych danych.

Implementacja transakcji poprzez zamki

- Zamek (lock): jedna z transakcji rezerwuje sobie dostęp do obiektu. Spójność przetwarzania transakcji jest zachowana, jeżeli cała transakcja ma dwie fazy: rosnącą i skracającą.

- Awaria w fazie potwierdzenia jest krytyczna, gdyż grozi utratą spójności
- W systemach rozproszonych faza potwierdzenia może trwać minuty i włącza zawodne elementy infrastruktury (łącza).
- Warunek atomowości oznacza, że jeżeli transakcja dzieli się na wiele podtransakcji wykonywanych na odległych serwerach, to nie może zajść sytuacja, gdy podtransakcja na serwerze A jest potwierdzona, zaś podtransakcja na serwerze B jest zerwana.
- Stąd konieczność dodatkowej ochrony fazy potwierdzenia, np. 2PC.

Protokoły potwierdzania

- W przypadku potwierdzania transakcji, wszystkie strony w niej uczestniczącej muszą przyjąć status jej wykonania
- Two-phase-commit protocol (2PC)
- Three-phase-commit protocol (3PC)

Two-Phase Commit - zasada działania

1. Transakcja została zakończona na wszystkich węzłach
2. Koordynator rozsyła do wszystkich węzłów biorących udział w transakcji zapytanie, czy ją akceptują. Każdy węzeł ją akceptuje lub odrzuca
3. Jeśli wszystkie węzły zaakceptowały transakcję, koordynator też ją akceptuje i powiadamia o tym węzły. Gdy któraś ze stacji odrzuci transakcję lub upłynie czas odpowiedzi, koordynator odrzuca transakcję i powiadamia o tym wszystkie węzły
4. W pewnych odmianach 2PC węzły dodatkowo przesyłają jeszcze jedno potwierdzenie (na wypadek gdyby któraś stacja zdążyła wysiąść)

Rozwiązywanie problemów dotyczących awarii

- Węzeł padnie, mimo że zaakceptował transakcję. W jego interesie leży, żeby dowiedział się (jak już wstanie) jak przebiegła akceptacja transakcji
- Koordynator przestaje działać – węzły czekają, aż się podniesie. Sytuacja niepożądana, bo awaria może się przedłużyć, dostęp do danych jest blokowany

Three-Phase Commit

- Działa podobnie do 2PC, istnieje jedna faza więcej
- Pierwsza faza identyczna z 2PC, druga jeśli zakończona sukcesem, wtedy koordynator wysyła prośbę potwierdzenia
- Jeśli co najmniej K stacji potwierdzi, transakcja jest akceptowana

Wybór koordynatora

- Koordynator nadzoruje zatwierdzenie / odrzucenie transakcji
- W przypadku gdy węzeł, na którym pracuje koordynator zawiedzie, powinien zacząć działać nowy koordynator na innym węźle, w przeciwnym wypadku dane transakcji są zablokowane
- Rozwiązanie: zapasowy koordynator lub algorytm elekcji

Zapasowy koordynator

- Działa równolegle z głównym koordynatorem, na innym węźle
- Odbiera wszystkie dane adresowane do głównego koordynatora,
- Nie podejmuje komunikacji z węzłami dopóki działa główny koordynator
- Utrzymuje połączenie z głównym koordynatorem, sprawdzając jego dyspozycyjność oraz synchronizując stan
- Jeśli łączność zostaje po jakimś czasie przerwana, uznaje, że główny koordynator nie działa, przejmuje jego zadania

Algorytm elekcji

- Każdy węzeł posiada unikatowy numer
- Koordynator nie odpowiada przez określoną ilość czasu ▫ trzeba wybrać nowy
- Węzeł zgłasza swoją kandydaturę na nowego koordynatora, rozsyła ją do wszystkich węzłów o numerze większym od numerze koordynatora, który przestał działać
- Jeśli nie dostaje odpowiedzi, oznacza to, że węzły nie działają, węzeł zostaje głównym koordynatorem, rozsyła tą informacje do wszystkich pozostałych węzłów, rozpoczyna zatwierdzanie transakcji od początku
- Jeśli dostaje odpowiedź, odczekuje na potwierdzenie, że inny węzeł będzie koordynatorem, jeśli potwierdzenie nie nadejdzie, algorytm jest restartowany
- W efekcie, węzeł o najwyższym numerze zostaje głównym koordynatorem

Kontrola współbieżności

- Single-Lock-Manager
- Multiple Coordinators
- Majority Protocol
- Biased Potocol
- Primary Copy

Single-Lock-Manager

- Manager znajduje się dokładnie w jednym miejscu w sieci
- Dane są blokowane na żądanie, jeśli manager nie może ich natychmiast zablokować, żądanie jest odkładane do momentu, w którym dane mogą zostać zablokowane
- W przypadku odczytu, dane mogą być pobrane z każdego serwera bazodanowego, zapis musi odbywać się na wszystkich węzłach
- Zalety
  - prostota implementacji,
  - łatwe usuwanie zakleszczeń,
- Wady
  - wąskie gardło,
  - niski poziom niezawodności

Multiple Coordinators

- Każdy węzeł posiada swojego managera
- Zasada działania podobnie jak w przypadku pojedynczego managera
- Zalety
  - większa przepustowość,
  - większa niezawodność (awaria pojedynczego węzła nie blokuje systemu)
- Wady
  - problem z zakleszczeniami

Majority Protocol

- W przypadku zapisu, należy zablokować każdą replikę
- Potwierdzenie blokady jest oczekiwane od każdego z węzłów zawierającego replikę
- Każda prośba blokady jest kolejkowana zgodnie z kolejnością przyjścia

Biased Potocol

- Rozróżniamy blokady typu Shared i Exclusive
- Shared - typowe dla odczytu, blokuje replikę na jednym z węzłów
- Exclusive - blokuje repliki na wszystkich węzłach, typowe dla zapisu
- Ten typ protokołu faworyzuje blokady typu Shared, działa podobnie do Majority, jest bardziej wydajny, gdyż odczyt występuje statystycznie częściej niż zapis

Primary Copy

- Jedna z replik jest uznawana za główną
- Blokada jest przetwarzana tylko na węźle głównej repliki
- Rzadziej występują zakleszczenia
- Jeśli węzeł z główną repliką padnie, dane stają się niedostępne, mimo, że fizycznie istnieją działające węzły z pozostałymi replikami

Zakleszczenia i jak sobie z nimi radzić (1)

- Zakleszczenie powstaje np. wtedy gdy węzeł A chce zablokować dane zablokowane przez węzeł B, a węzeł B chce zablokować dane blokowane przez węzeł A
- W/w sytuację można rozszerzyć na większą ilość węzłów
- W celu określenia zapotrzebowania na dane możemy stworzyć graf, którego węzłami są transakcje, krawędziami zaś zapotrzebowanie na dostęp do danych
- Na każdym węźle fizycznym istnieje graf, który reprezentuje bieżącą sytuację

Zakleszczenia i jak sobie z nimi radzić (2)

- Na każdym węźle fizycznym rezyduje kontroler, który buduje graf, uzupełnia go, wyszukuje zakleszczenia
- Każdy kontroler ma graf lokalny, na jednym węźle budowany jest graf globalny
- Jeśli kontroler wykryje zakleszczenie lokalne (na swoim węźle) usuwa je
- Usunięcie zakleszczenia polega na cofnięciu transakcji jednego z węzłów-ofiary, do momentu odblokowania zakleszczenia. Transakcja ta jest wznawiana, gdy pozostałe transakcje, biorące udział w zakleszczeniu zostaną zakończone
- Kontroler wysyła nową postać grafu do węzła z kompletnym grafem zakleszczeń

Pytania?