Środowiska przetwarzania rozproszonego

Dariusz Wawrzyniak

✉ Politechnika Poznańska
Instytut Informatyki
ul. Piotrowo 2 (IIPP, pok. 5)
60-965 Poznań

Dariusz.Wawrzyniak@cs.put.poznan.pl

www.cs.put.poznan.pl/dwawrzyniak

☎ +48 (61) 665 2963

Program przedmiotu

- Wykład (16 godz.):
  - założenia projektowe w budowie systemów rozproszonych
  - transparentność dostępu do zdalnych zasobów (modele komunikacji)
  - skalowalność (replikacja)
- Ćwiczenia laboratoryjne (16 godz.)
  implementacja przykładowych aplikacji rozproszonych z wykorzystaniem wybranych mechanizmów/środowisk:
  - PVM/MPI (wymiana komunikatów)
  - RPC (wywoływanie procedur zdalnych)
  - CORBA (podejście obiektowe —wywoływanie metod zdalnych)

Sposób zaliczenia przedmiotu

- Wykład (16 godz.)
  egzamin końcowy.
- Ćwiczenia laboratoryjne (16 godz.)
  testy/sprawdziany z zagadnień laboratoryjnych w trakcie ćwiczeń (ewentualnie dodatkowo w ramach egzaminu końcowego).

Literatura

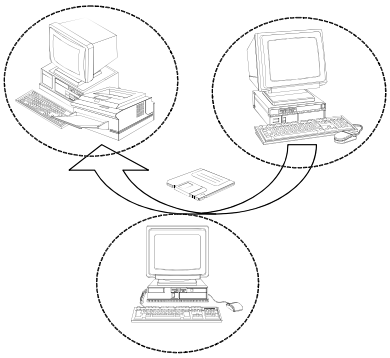
- A. S. Tanenbaum, M. van Steen: Systemy rozproszone, Zasady i paradygmaty. WNT, Warszawa, 2006.
- G. Coulouris, J. Dollimore, T. Kindberg: Systemy rozproszone: podstawy i projektowanie. WNT, Warszawa, 1998.
- A. S. Tanenbaum: Rozproszone systemy operacyjne, PWN, Warszawa, 1997.

System rozproszony od strony projektowej

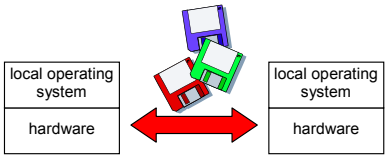
System rozproszony

- Zbiór niezależnych komputerów i zasobów komputerowych zdolnych do kooperacji (np. poprzez sieć komputerową), postrzeganych przez użytkownika jako całościowo spójny system.
- Ogólny cel projektowy systemu rozproszonego: stworzenie przezroczystego, otwartego, elastycznego, wydajnego i skalowalnego mechanizmu współdzielenia zasobów.

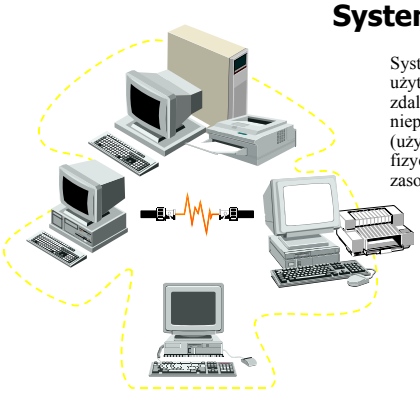
Niezależne komputery



Niezależne komputery — organizacja oprogramowania

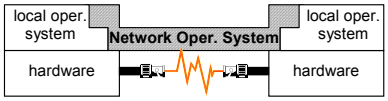


System sieciowy

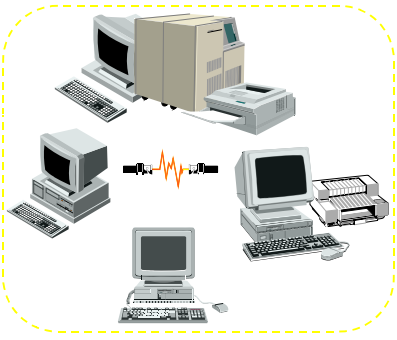


System sieciowy umożliwia użytkownikom dostęp do zdalnych zasobów w sposób nieprzezroczysty (użytkownicy są świadomi fizycznego rozproszenia zasobów)

System sieciowy — organizacja oprogramowania

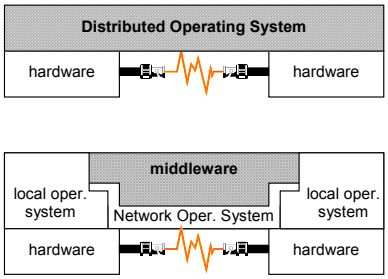


System rozproszony



Rozproszony system operacyjny umożliwia użytkownikom dostęp do zdalnych zasobów w sposób przezroczysty (jest jednolity sposób dostępu do zasobu lokalnego i zdalnego)

System rozproszony — organizacja oprogramowania



Zagadnienia projektowe (1)

- Otwartość, elastyczność — zdolność do rozbudowy (sprzętowej, programowej) i rekonfiguracji, możliwość dodawania nowych usług bez głębokiej ingerencji w istniejące już usługi.
- Skalowalność — możliwość dostosowania systemu do rosnących rozmiarów problemów i wymagań użytkowników.
- Współbieżność — możliwość współbieżnego (równoczesnego) ubiegania się o zasoby i ich użytkowania.

Zagadnienia projektowe (2)

- Tolerowanie uszkodzeń — odporność na awarie przez redundancję (sprzętu, danych) oraz możliwość odtworzenia spójnego stanu po awarii.
- Przezroczystość — ukrywanie przed użytkownikiem (programistą) fizycznego odseparowania składowych, zapewnienie jednolitego dostępu do zasobów lokalnych i zdalnych.
- Wydajność — optymalizacja ruchu w sieci w celu zredukowania negatywnego wpływy stosunkowo wolnej komunikacji sieciowej.

Przezroczystość (1)

- przezroczystość dostępu — ukrywanie różnic w reprezentacji danych, zagwarantowanie jednolitego sposobu dostępu do zasobów, niezależnie od tego, czy są to zasoby lokalne, czy zdalne
- przezroczystość położenia — identyfikacja zasobów niezależna od ich fizycznej lokalizacji (np. usługa nazw)
- przezroczystość migracji — zmiana fizycznej lokalizacji zasobu nie powoduje zmian w sposobie ich identyfikacji i w sposobie dostępu
- przezroczystość relokacji — fizyczna zmiana lokalizacji zasobu może być dokonana w sposób niewidoczny dla aplikacji w czasie realizacji dostępu do niego przez użytkowników

Przezroczystość (2)

- przezroczystość replikacji — utrzymywanie i udostępnianie kilku egzemplarzy tego samego zasobu (kopii) w taki sposób, jak gdyby użytkownik widział i działał tylko na jednym
- przezroczystość awarii — zdolność ukrycia przed użytkownikiem faktu chwilowych nieprawidłowości funkcjonowania
- przezroczystość współbieżności — realizacja współbieżnego dostępu do zasobu w taki sposób, że konkurujące procesy nie przeszkadzają sobie wzajemnie

Otwartość

- standaryzacja reguł dostępu do usług — formalny opis składni (jak skorzystać z usługi) i semantyki (na czym polega realizacja usługi)
- interoperacyjność (interoperability) — zdolność współpracy dwóch różnych systemów, korzystający wzajemnie jedynie ze swoich własnych usług
- przenośność (portablility) — zdolność aplikacji zaprojektowanej w systemie rozproszonym A do działania bez modyfikacji w systemie B

Polityka i mechanizm

- Polityka jest zbiorem reguł dostępu do zasobów i może wynikać z:
  - projektu systemu (ustalona jest na etapie projektowania systemu)
  - wytycznych kierownictwa (ustalana jest na etapie instalacji lub eksploatacji systemu)
  - decyzji indywidualnych użytkowników (podjętych w trakcie eksploatacji systemu)
- Mechanizm jest zbiorem dostępnych środków do wymuszania polityki
  - mechanizm powinien być na tyle uniwersalny, żeby dało się go dostosować do zmian w polityce ➔ **otwartość**
  - w skrajnym przypadku każda zmiany polityki mogłaby wymagać zmiany mechanizmu ➔ **system zamknięty**

Koncepcja dostępu do współdzielonych zasobów

- zarządca zasobu (ang. resource manager) — moduł oprogramowania odpowiedzialny za udostępnianie zasobu
- użytkownik zasobu — moduł (proces) zgłaszający zapotrzebowanie na zasób

Współpraca pomiędzy zarządcą a użytkownikiem

- model komunikatowy — współpraca odbywa się przez komplementarne wykonanie operacji *send* i *receive* odpowiednio przez nadawcę i odbiorcę komunikatu
- model obiektowy — zarządca postrzegany jest jako obiekt o pewnym identyfikatorze, będący w określonym stanie, który zmienia się pod wpływem operacji żądanych przez użytkowników
  - obiekt aktywny — obiekt jest procesem (wielowątkowym), oczekującym na żądania wywołania metod
  - obiekt pasywny — każdy obiekt ma własną przestrzeń adresową (segment), odwzorowywaną na przestrzeń adresową procesu, który się do niego odwołuje

Komunikacja w systemach sieciowych/rozproszonych

- Elementarne mechanizmy komunikacji pomiędzy procesami
- Model komunikacji w systemach rozproszonych

Elementarne mechanizmy komunikacji pomiędzy procesami

- Współdzielenie pamięci
  - podstawą komunikacji jest dostęp do wspólnych danych w pamięci,
  - wymiana informacji sprowadza się do zapisu i odczytu wspólnych danych oraz związanej z tym synchronizacji.
- Przekazywanie komunikatów
  - podstawą komunikacji jest umieszczanie danych w podsystemie komunikacyjnym oraz ich pobieranie z podsystemu komunikacyjnego,
  - wymiana informacji polega na wywoływaniu odpowiednich funkcji w celu wysłania i odbioru komunikatu.

Sieciowa realizacja elementarnych mechanizmów komunikacji

- Współdzielenie pamięci
  - brak możliwości współdzielenia pamięci fizycznej,
  - emulacja pamięci współdzielonej poprzez odpowiednią obsługę błędu strony w systemie pamięci wirtualnej.
- Przekazywanie komunikatów
  - komunikacja asynchroniczna → wymagana gotowość do odbioru danych po stronie adresata komunikatu,
  - schemat komunikacji zgodny z modelem klient-serwer.

Model komunikacji w systemach rozproszonych

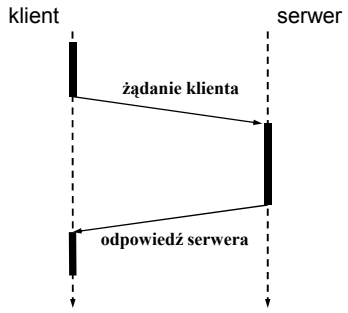
- Wywoływanie procedur zdalnych (ang. remote procedure call)
- Wywoływanie metod zdalnych (ang. remote method call)
- Komunikacja zorientowana na przysyłanie wiadomości (ang. message-oriented communication)
- Komunikacja strumieniowa
- Komunikacja za pośrednictwem wirtualnej pamięci współdzielonej

Wywoływanie procedur zdalnych

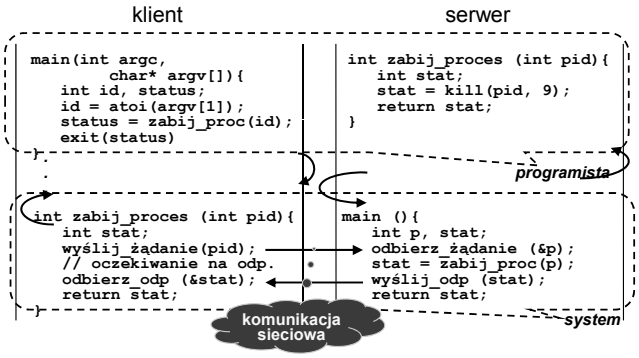
Mechanizm wywołania procedury

```
main(int argc, char* argv[]){
   int id, status;
   id = atoi(argv[1]);
   status = zabij_proc(id);
   exit(status)
}  .
   .
   .
int zabij_proces (int pid){
   int stat;
   stat = kill(pid, 9);
   return stat;
}
```

Schemat interakcji klient-serwer (transakcja komunikatu)



Semantyka wywołania procedury w komunikacji sieciowej



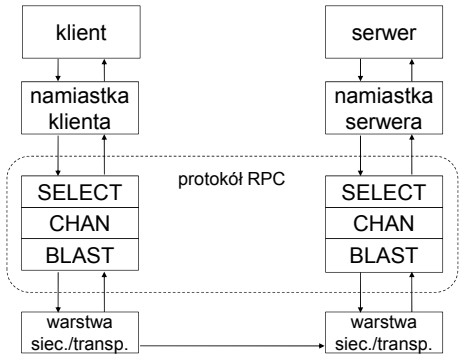
RPC — zagadnienia projektowe

- Przeźroczystość dostępu — ukrycie komunikacji sieciowej przed aplikacją przez odpowiednie opakowanie funkcji komunikacyjnych namiastką klienta oraz serwera.
- Gwarancja wykonania — ukrywanie błędów komunikacyjnych
- Specyfikacja interfejsu — sposób opisu sygnatur procedur zdalnych (nazwy, typy parametrów)
- Obsługa sytuacji wyjątkowych

RPC — przezroczystość dostępu

- Namiastka klienta (ang. client stub) — udostępnienie aplikacji klienckiej procedury lokalnej odpowiedzialnej za przesłanie danych do serwera oraz odebranie wyników
- Namiastka serwera (ang. server stub) — udostępnienie aplikacji po stronie serwera procedury lokalnej odpowiedzialnej za odebranie identyfikatora procedury zdalnej do wywołania, parametrów procedury, a odesłanie wyników lub zgłoszenie wyjątków

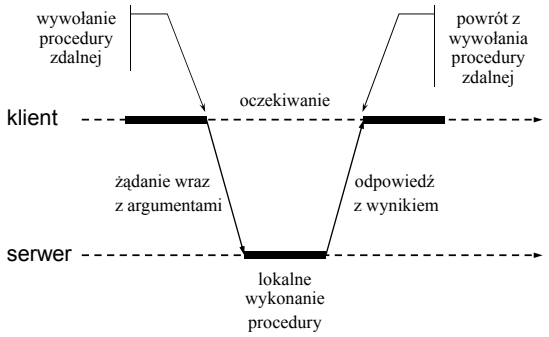
Warstwy protokołu RPC



Interakcja klient-serwer w realizacji wywołania zdalnego

1. Lokalne wywołanie procedury namiastki przez klienta
2. Przygotowanie (upakowanie) danych do wysłania na stronę serwera
3. Wysłanie przygotowanego komunikatu na stronę serwera
4. Odebranie komunikatu przez serwer
5. Rozpakowanie danych
6. Wykonanie procedury zdalnej
7. Przygotowanie (upakowanie) danych z odpowiedzią dla klienta
8. Wysłanie przygotowanego komunikatu na stronę klienta
9. Odebranie komunikatu przez namiastkę klienta
10. Rozpakowanie komunikatu i zwrócenie klientowi wyniku

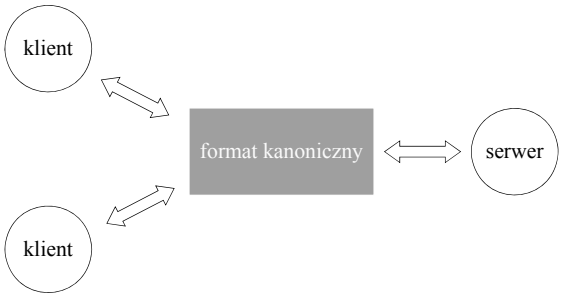
Diagram przestrzenno-czasowy interakcji klient-serwer



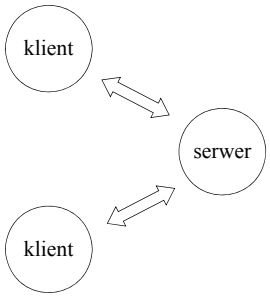
RPC - przekazywanie parametrów

- przekazywanie przez wartość (ang. call-by-value) — problem reprezentacji danych (np. różnice w kodowaniu znaków, różnice w kolejności bajtów, formaty liczb zmiennopozycyjnych)
- przekazywanie przez referencje (ang. call-by-reference) — problem zinterpretowania wartości wskaźnika w innej przestrzeni adresowej
- przekazywanie przez kopiowanie i odtwarzanie (ang. call-by-copy/restore) — utworzenie kopii parametru po stronie procedury i zapis dokonanych tam zmian w procesie klienta po jej zakończeniu (problem opisu struktur danych w celu prawidłowego zidentyfikowania wszystkich składowych)

Reprezentacja kanoniczna



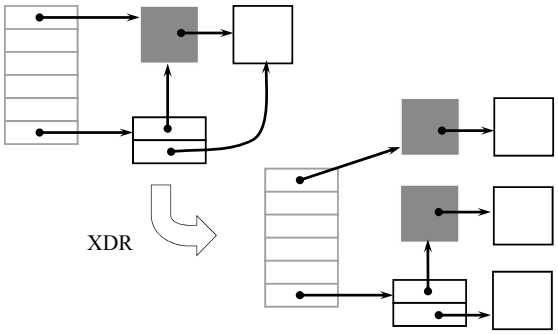
Konwersja bezpośrednia



Przekazywanie referencji przez kopiowanie i odtwarzanie

1. Skopiowanie wskazanej wartości do bufora komunikacyjnego i wysłanie do serwera.
2. Wywołanie po stronie serwera procedury zdalnej ze wskaźnikiem na kopię wartości utworzoną po stronie serwera.
3. Skopiowanie zmodyfikowanej wartości z przestrzeni adresowej serwera do bufora komunikacyjnego i przesłanie z powrotem do klienta.
4. Umieszczenie odebranej wartości w miejscu wskazywanym przez referencję po stronie klienta.

Kopiowanie i odtwarzanie — przykład złożonej struktury danych



RPC — gwarancja wykonania

- Semantyka *ewentualnie* — brak gwarancji, procedura mogła się wykonać lub mogła się nie wykonać.
- Semantyka *co najmniej raz* — po uzyskaniu **odpowiedzi z wynikiem** od serwera klient ma pewność, że wywoływana procedura wykonała się co najmniej raz.
- Semantyka *co najwyżej raz* — po uzyskaniu **odpowiedzi z wynikiem** od serwera klient wie, że wywoływana procedura wykonała się dokładnie raz.

Jak należy zinterpretować przypadek wystąpienia błędu (wyjątku) w wywołaniu procedury zdalnej?

- Semantyka *dokładnie raz* — niemożliwa do uzyskania, jeśli system narażony jest na awarie (np. serwera lub łączy).

Specyfikacja interfejsu

- Opis interfejsu w języku implementacji (np. Ada, Java RMI)
- Opis interfejsu w języku specjalnym, niezależnym od implementacji (CORBA — IDL, Sun RPC — rpcgen)

RPC — zagadnienia realizacyjne

- Przetwarzanie interfejsu
- Wiązanie klienta z serwerem
- Obsługa komunikacji klient-serwer
- Realizacja semantyki błędu
- Problem osieroconych obliczeń

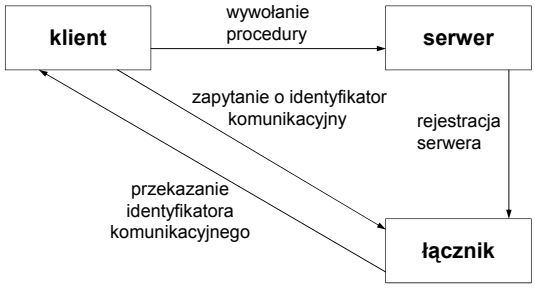
Przetwarzanie interfejsu procedur zdalnych

- Generowanie namiastki klienta
- Generowanie namiastki serwera
- Generowanie przykładowego programu klienta (client sample)
- Generowanie wzorca do implementacji procedur zdalnych (template)
- Generowanie plików do zarządzania kompilacją

Wiązanie klienta z serwerem

- Wiązanie statyczne — klient ma na stałe wprowadzony identyfikator komunikacyjny serwera (np. para: adres IP, nr portu).
- Wiązanie dynamiczne — klient uzyskuje adres serwera za pośrednictwem łącznika (np. portmap, lub rpcbind w Sun RPC).

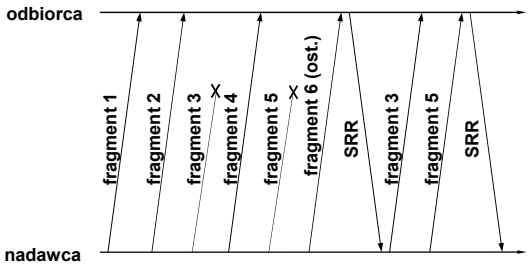
Wiązanie dynamiczne klienta z serwerem



Obsługa komunikacji klient-serwer

- BLAST — realizuje przesyłanie dużych komunikatów poprzez podział na mniejsze części, transmisję poszczególnych części i ponowne złożenie w jeden komunikat po stronie odbiorczej,
- CHAN — synchronizuje wymianę komunikatów z żądaniami wywołania procedur oraz odpowiedziami,
- SELECT — rozdziela i przekazuje komunikaty z żądaniami do odpowiednich procesów.

BLAST



BLAST — nadawca

1. otrzymanie bloku z w wyższej warstwy stosu protokołów i podział bloku na fragmenty
2. wysłanie kolejno poszczególnych fragmentów do odbiorcy (ze specjalnym oznaczeniem ostatniego fragmentu)
3. ustawienie czasomierza (ang. timer) DONE
4. jeśli dotarł SRR z informacją o brakujących fragmentach, to wysłanie brakujących fragmentów i przejście do pkt. 3
5. jeśli dotarł SRR z informacją o odebraniu wszystkich fragmentów,
   to SUKCES
6. jeśli DONE = 0 // minął czas oczekiwania
   to BŁĄD

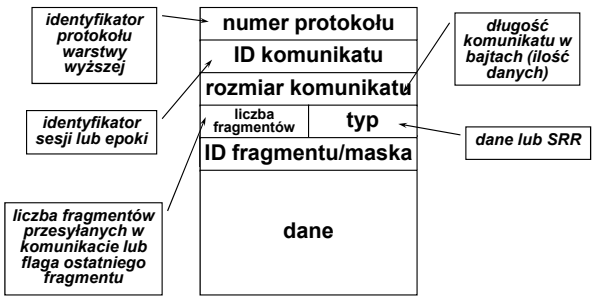
BLAST — odbiorca (1)

1. po odebraniu pierwszego komunikatu z fragmentem bloku danych:
   inicjalizacja struktur danych do przechowywania poszczególnych bloków, umieszczenie odebranego fragmentu w odpowiedniej strukturze
   ustawienie licznika powtórzeń na 0
2. ustawienie czasomierza LAST_FRAG
3. po odebraniu kolejnego (ale nie ostatniego) komunikatu: umieszczenie nadesłanego fragmentu bloku w odpowiedniej strukturze i przejście do pkt. 2

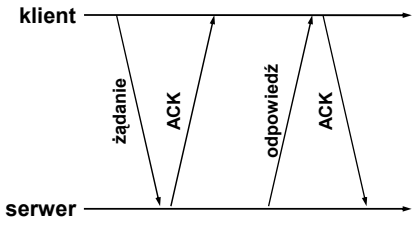
BLAST — odbiorca (2)

4. jeśli LAST_FRAG = 0, to przejście do pkt. 8
5. po odebraniu ostatniego komunikatu umieszczenie nadesłanego fragmentu bloku w odpowiedniej strukturze i sprawdzenie kompletności bloku
6. jeśli blok jest kompletny to wysłanie komunikatu SRR i przekazanie bloku do wyższej warstwy stosu protokołów
7. jeśli blok nie jest kompletny przejście do pkt. 8
8. jeśli licznik powtórzeń jest mniejszy od 3, to wysłanie komunikatu SRR z informacją o brakujących blokach, zwiększenie licznika powtórzeń o 1 i przejście do pkt. 2 w przeciwnym razie rezygnacja z odbioru

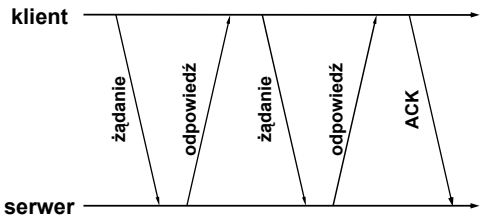
BLAST — format komunikatu



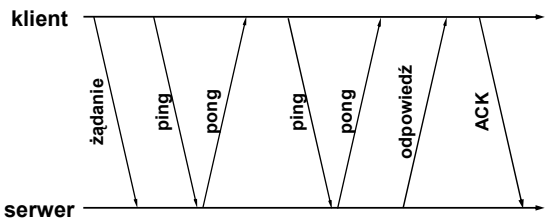
CHAN



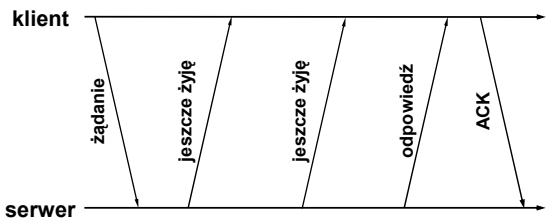
CHAN — domniemane potwierdzenia



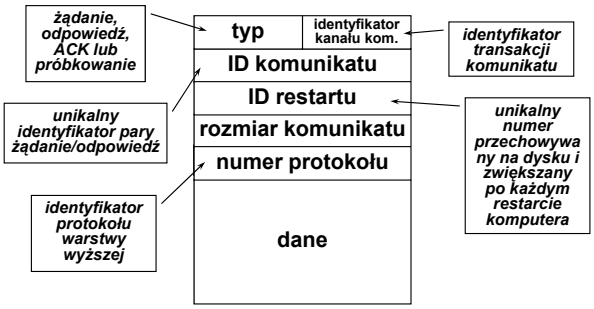
CHAN — próbkowanie serwera



CHAN — „bicie serca”



CHAN — format komunikatu



Realizacja semantyki błędu

- nie można zlokalizować serwera — zgłoszenie wyjątku
- zaginione żądanie — retransmisja żądania po upłynięciu ustalonego czasu oczekiwania
- zaginiona odpowiedź — retransmisja żądania
  - stosowanie procedur idempotentnych
  - numerowanie żądań i retransmisja odpowiedzi
- awaria serwera
  - przed podjęciem realizacji → retransmisja żądania
  - po wykonaniu →zgłoszenie wyjątku
- awaria klienta — osierocenie obliczeń

Usuwanie osieroconych obliczeń (1)

- Eksterminacja — rejestrowanie działań podejmowanych przez klienta na nośniku niewrażliwym na awarie i usuwanie na tej podstawie osieroconych obliczeń po restarcie klienta.
- Reinkarnacja — każdy restart klienta rozpoczyna nową epokę (identyfikowaną przez numer kolejny), po której usuwane są wszystkie obliczenia związane z poprzednią epoką.

Usuwanie osieroconych obliczeń (2)

- Łagodna reinkarnacja — reinkarnacja, w której usuwa się tylko te obliczenia rozpoczęte w starej epoce, dla których nie ma właściciela.
- Wygaśnięcie — przydział określonego czasu *T* serwerowi na wykonanie procedury. Jeśli wykonanie nie zakończy się w czasie *T*, serwer musi uzyskać kolejny przydział, pod warunkiem, że obliczenia nie zostały osierocone. Jeśli klient odczeka czas *T* przy restarcie, osierocone obliczenia same się zakończą.

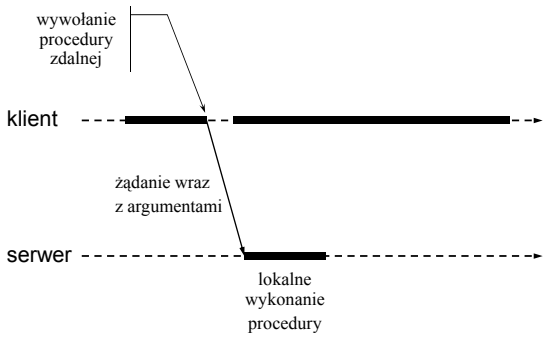
SELECT

- Po stronie klienta: odwzorowanie wywoływanej procedury na jej identyfikator, przekazywany do serwera.
- Po stronie serwera: zlokalizowanie wywoływanej procedury na podstawie identyfikatora.

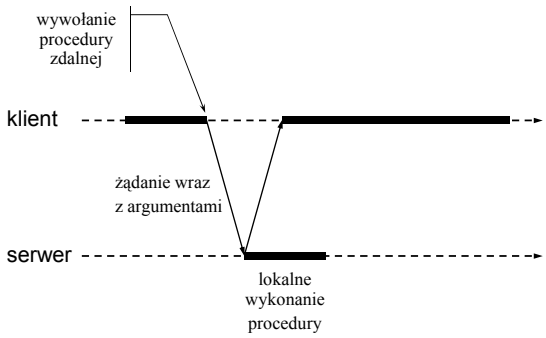
Warianty użycia mechanizmu RPC

- Wywołanie asynchroniczne
- Wywołanie zwrotne

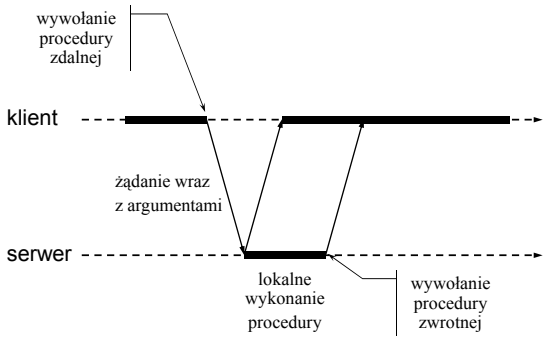
Wywołanie asynchroniczne



Wywołanie asynchroniczne z potwierdzeniem odbioru



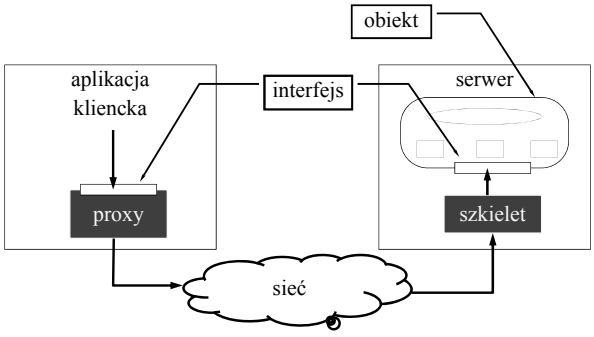
Wywołanie zwrotne



Wywoływanie metod zdalnych

Podejście obiektowe do budowy systemów rozproszonych

Wywoływanie metod zdalnych — model systemu



Istota podejścia obiektowego

- Obiekt jest jednostką integrującą w sobie dane (stan obiektu) oraz odpowiednio zdefiniowane operacje (metody)
- Jedyna forma dostępu do danych polega na wywoływaniu metod wyspecyfikowanych w publicznym interfejsie obiektu
- Obiekt może implementować wiele interfejsów
- Ten sam interfejs może być implementowany przez wiele obiektów

Podejście obiektowe do budowy systemów rozproszonych

- Kluczowe dla mechanizmu wywoływania metod zdalnych jest oddzielenie definicji interfejsu (specyfikacji) od jego implementacji w obiekcie
- W językach definicji interfejsu (IDL) interfejs traktowany jest jak typ danych
- W języku implementacji każdy obiekt implementujący dany interfejs jest instancją tego typu

Przykład opisu interfejsu w podejściu obiektowym (CORBA IDL)

```
interface Konto {
     float stan();
     float wplac(in float value);
     float pobierz(in float value);
};

interface Bank {
     Konto otworzKonto(in int numer);
     float zamknijKonto(in Konto k);
};
```

Klasyfikacja obiektów ze względu na sposób implementacji

- Obiekt zdalny (ang. remote object) — stan obiektu utrzymywany jest przez jeden serwer (wszystkie dane obiektu znajdują się na jednym serwerze)
- Obiekt rozproszony (ang. distributed object) — stan obiektu przechowywany jest przez więcej niż jeden serwer (poszczególne serwery przechowują odpowiednie dane, stanowiące fragmenty tego samego obiektu)
- Z punktu widzenie klienta zarówno obiekt zdalny jak i obiekt rozproszony stanowi pewną całość odpowiednio identyfikowaną

Klasyfikacja ze względu na dostępność informacji o obiektach

- Obiekt dostępny w czasie kompilacji (ang. compile-time object) — informacja o obiekcie (jego typ) znana jest i dostępna w odpowiedniej formie w programie klienta na etapie tworzenia oprogramowania
- Obiekt dostępny w czasie wykonania (ang. runtime object) — informacja o obiekcie dostępna jest dopiero w czasie działania aplikacji

Klasyfikacja obiektu ze względu na trwałość

- Obiekt trwały (ang. persistent object) — obiekt istnieje nawet wówczas, gdy nie ma serwera, w przestrzeni adresowej, którego mógłby być przechowywany stan tego obiektu
- Obiekt przejściowy (ang. transient object) — obiekt istnieje tak długo, jak długo działa serwer utrzymujący jego stan

Referencja do obiektu

- Referencja jest identyfikatorem, wykorzystywanym do wskazania obiektu, na którym ma zostać wykonana operacja
- W celu wywołania metody obiektu rezydującego na innej maszynie potrzebna jest zdalna referencja
- Referencja może być przekazywana jako parametr wywołania metody lub zwrócona jako wynik wykonania metody

Implementacja zdalnej referencji

- Referencja jest wartością, której struktura wewnętrzna nie jest w żaden sposób interpretowana przez aplikację
- Zdalna referencja musi zawierać wystarczająco dużo informacji, żeby dowiązać obiekt w programie klienta
- Dowiązanie obiektu oznacza uzyskanie informacji, niezbędnej do wywołania metody obiektu (uzyskanie proxy)

Dowiązanie obiektu

- Dowiązanie jawne (ang. explicit binding) — dowiązanie wymaga wywołania odpowiedniej funkcji (np. bind), po którym może dopiero nastąpić wywołanie metody
- Dowiązanie domyślne (ang. implicit binding) — wywołanie metody za pośrednictwem referencji skutkuje ustanowieniem dowiązanie do obiektu w procesie klienta

Wywoływanie metod

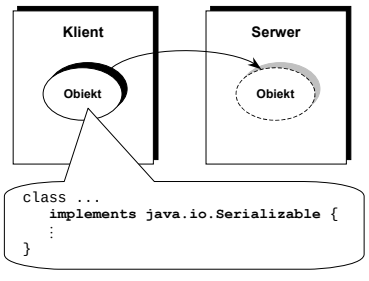
- Statyczne wywoływanie metod — wywołanie metody za pośrednictwem proxy, wygenerowanego na podstawie definicji interfejsu
- Dynamiczne wywoływanie metod — „skomponowanie" informacji niezbędnych do wywołania w czasie działania systemu, np.:

```
invoke(ref_obj, id_metody,
       param_wej, param_wyj);
```

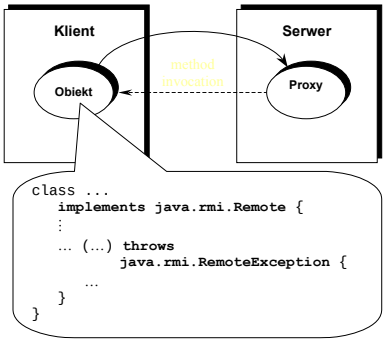
Przekazywanie obiektów jako parametrów

- Przekazywanie przez wartość — do zdalnej metody przekazywana jest kopia obiektu, który jest parametrem rzeczywistym wywołania.
- Przekazywanie przez referencję (zmienną) — do zdalnej metody przekazywana jest referencja (zdalna) do obiektu, który jest parametrem rzeczywistym.
- Przekazywanie przez kopiowanie i odtwarzanie — do zdalnej metody przekazywana jest kopia obiektu, a po zakończeniu zdalnej metody następuje aktualizacja obiektu po stronie wywołania, stosownie do zmian dokonanych w ramach wykonania metody.

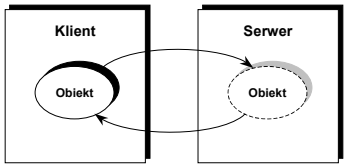
Przekazywanie przez wartość — przykład w Java RMI



Przekazywanie przez referencję — przykład w Java RMI

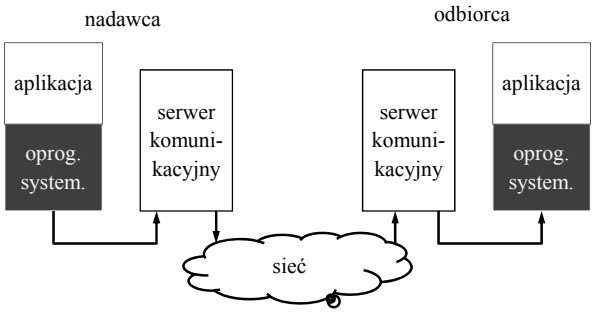


Przekazywanie przez kopiowanie i odtwarzanie



Komunikacja zorientowana na przysyłanie wiadomości

Przesyłanie wiadomości — model systemu



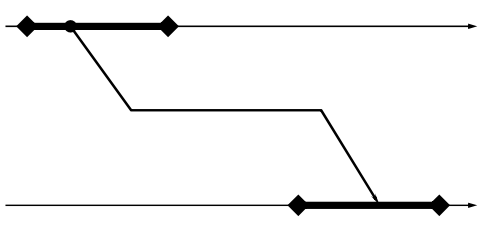
Trwałość komunikacji

- Komunikacja przejściowa (ang. transient communication) — wiadomość jest przekazywana (utrzymywana w podsystemie komunikacyjnym) pod warunkiem, że działa nadawca i odbiorca tej wiadomości.
- Komunikacja nieustanna (ang. persistent communication) — wiadomość jest przechowywana w celu doręczenia do odbiorcy nawet, gdy odbiorca nie działa w danej chwili, a nadawca zakończył działanie po wysłaniu tej wiadomości.

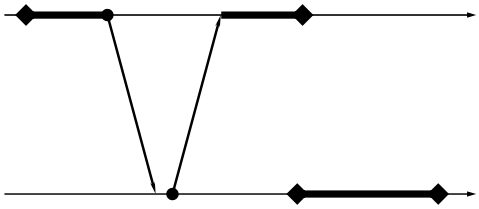
Synchroniczność komunikacji

- Komunikacja synchroniczna — nadawca kontynuuje działanie dopiero gdy wiadomość znajdzie się w buforze odbiorczym lub zostanie doręczona do adresata.
- Komunikacja asynchroniczna — nadawca kontynuuje działanie zaraz po przekazaniu wiadomości do podsystemu komunikacyjnego.

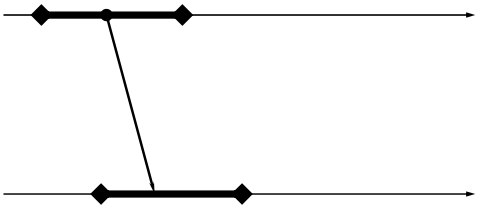
Komunikacja nieustanna asynchroniczna



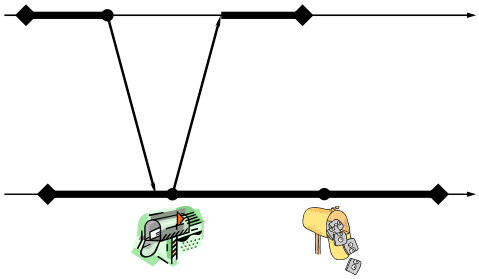
Komunikacja nieustanna synchroniczna



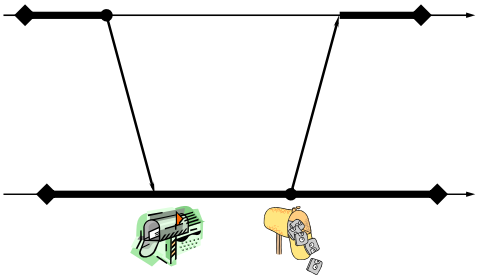
Komunikacja przejściowa asynchroniczna



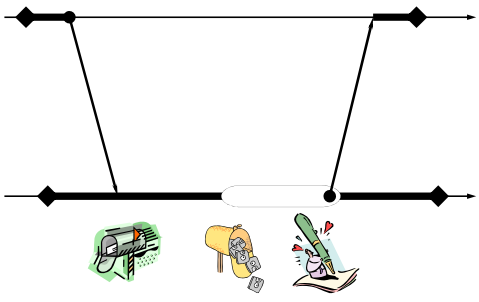
Komunikacja przejściowa synchroniczna z potwierdzeniem dotarcia



Komunikacja przejściowa synchroniczna z potwierdzeniem odebrania



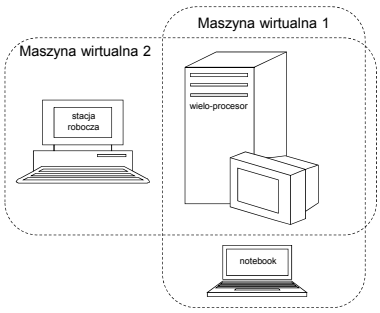
Komunikacja przejściowa synchroniczna z oczekiwaniem na odpowiedź



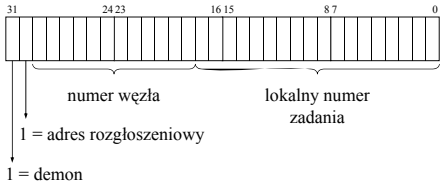
Środowiska wymiany komunikatów

- Identyfikacja procesów — przeźroczystość położenia
  - PVM: unikalny identyfikator procesu (TID)
  - MPI: unikalny numer w grupie procesów
- Mechanizmy komunikacji — przeźroczystość dostępu
  - PVM: adresowanie komunikatów do procesów o podanych identyfikatorach
  - MPI: adresowanie komunikatów do procesów o podanych numerach
- Mechanizm zdalnego uruchamiania zadań
  - PVM: dynamicznie w trakcie działania aplikacji
  - MPI (v.1): statycznie w trakcie uruchamiania aplikacji

PVM — maszyna wirtualna



PVM — budowa identyfikatora zadania (TID)



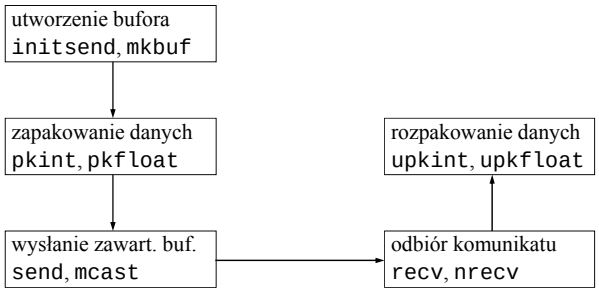
PVM — uruchamianie przetwarzania

- Przygotowanie programów
  - przygotowanie kodów źródłowych w języku C lub Fortran
  - kompilacja i konsolidacja kodów źródłowych
- Skonfigurowanie maszyny wirtualnej
  - wybranie i przygotowanie opisu odpowiednich węzłów
  - uruchomienie demonów na poszczególnych węzłach
  - ewentualna dynamiczna zmiana konfiguracji początkowej
- Uruchomienie zadań
  - uruchomienie procesu na terminalu jednego z węzłów
  - uruchomienie zadania poleceniem `spawn` z konsoli PVM
  - uruchomienie zadania przez inne zadanie

PVM — podstawowe funkcje biblioteczne

- Konfiguracja maszyny wirtualnej: `pvm_addhosts`, `pvm_delhosts`, `pvm_config`, `pvm_tidtohost`
- Obsługa zadań: `pvm_mytid`, `pvm_exit`, `pvm_spawn`, `pvm_kill`, `pvm_task`, `pvm_parent`
- Komunikacja międzyprocesowa
  - obsługa buforów: `pvm_initsend`, `pvm_mkbuf`, `pvm_freebuf`, `pvm_getsbuf`, `pvm_getrbuf`, `pvm_setsbuf`, `pvm_setrbuf`
  - pakowanie danych: `pvm_pk`..., `pvm_upk`...
  - wymiana komunikatów: `pvm_send`, `pvm_mcast`, `pvm_psend`, `pvm_recv`, `pvm_nrecv`, `pvm_probe`, `pvm_trecv`, `pvm_bufinfo`, `pvm_precv`

PVM — schemat wymiany komunikatów



PVM — dynamiczne grupy procesów

- Grupa procesów identyfikowana jest przez nazwę.
- Każde zadanie może w dowolnej chwili dołączyć się do grupy, jak i opuścić grupę.
- Przyłączając się do grupy, zadanie otrzymuje w tej grupie unikalnym numer (jest to numer kolejny, począwszy od 0).
- Zadanie może należeć jednocześnie do wielu grup.
- Zadanie może wysłać komunikat do wszystkich procesów w grupie, nawet jeśli do niej nie należy (grupy otwarte).

PVM — funkcje biblioteczne do obsługi grup procesów

- Dołączanie procesu do grupy (również tworzenie grupy): `pvm_joingroup`
- Odłączanie procesu od grupy (ostatecznie usuwanie grupy): `pvm_lvgroup`
- Identyfikacja procesów w grupie: `pvm_gettid`, `pvm_getinst`
- Informacja o liczbie procesów w grupie: `pvm_gsize`
- Rozgłaszanie (komunikat do wszystkich w grupie): `pvm_bcast`
- Bariera synchronizująca: `pvm_barrier`

MPI — podstawowe cechy

- Model przetwarzania — SPMD (Single Program Multiple Data)
  - wszystkie uruchomione procesy wykonują ten sam program
- Wszystkie procesy uruchamiane są przy rozpoczęciu przetwarzania (w wersji 2.0 można również dynamicznie uruchomić dodatkowe procesy) — `mpirun`, `mpiexec`
- Procesy tworzą grupę, w której numer procesu jest jego identyfikatorem
  - grupa jest częścią tzw. komunikatora
  - istnieje predefiniowany komunikator `MPI_COMM_WORLD` obejmujący wszystkie procesy

MPI — komunikator

- Kontekst komunikacyjny — wirtualny kanał komunikacyjny, umożliwiający odseparowanie komunikatów
  - można odbierać tylko komunikaty przekazywane w ramach tego samego kontekstu
- Grupa procesów — grupa o ustalonym rozmiarze, w ramach której identyfikowane są procesy
  - nadawca i odbiorca identyfikowany jest poprzez numer w danej grupie
- Komunikator jest parametrem każdej funkcji do realizacji wymiany komunikatów

MPI — obsługa komunikatora

- Uzyskanie własnego numeru przez proces: `MPI_Comm_rank`
- Uzyskanie liczby procesów w grupie komunikatora: `MPI_Comm_size`
- Uzyskanie grupy komunikatora: `MPI_Comm_group`
- Utworzenie komunikatora dla określonej grupy procesów: `MPI_Comm_create`

MPI — komunikacja punkt-punkt

- Komunikacja w trybie blokującym: `MPI_Send`, `MPI_Recv`
- Komunikacja natychmiastowa: `MPI_Isend`, `MPI_Irecv`
- Komunikacja synchroniczna: `MPI_Ssend`, `MPI_Issend`
- Komunikacja buforowana: `MPI_Bsend`, `MPI_Ibsend`
- Komunikacja w trybie gotowości: `MPI_Rsend`, `MPI_Irsend`

MPI — przetwarzanie kolektywne

- Kolektywny transfer danych: `MPI_Bcast`, `MPI_Gather`, `MPI_Scatter`, `MPI_Gatherall`, `MPI_Alltoall`
- Obliczenie kolektywne: `MPI_Reduce`, `MPI_Reducescatter`
- Synchronizacja: `MPI_Barrier`

MPI_Alltoall

A1	A2	A3	A4
B1	B2	B3	B4
C1	C2	C3	C4
D1	D2	D3	D4

A1	B1	C1	D1
A2	B2	C2	D2
A3	B3	C3	D3
A4	B4	C4	D4

Message Oriented Middleware

Systemy kolejkowania komunikatów

Cechy MOM

- Uniezależnienie funkcjonowania składników aplikacji od dostępności informacji o interfejsach innych składników
- Uniezależnienie funkcjonowania warstwy komunikacyjnej (kanału komunikacyjnego) od działania (obecności) komunikujących się procesów
- Łatwość wdrożenia komunikacji asynchronicznej

Koncepcja komunikacji oparta na kolejkowaniu

- Organizacja warstwy komunikacyjnej w postaci systemu kolejek realizowanych w oparciu o zasoby pamięci, w tym pamięci dyskowej (gwarancja trwałości na wypadek awarii)
- Udostępnienie mechanizmów komunikacji polegających na:
  - umieszczeniu komunikatów w kolejkach,
  - pobieraniu komunikatów z kolejek

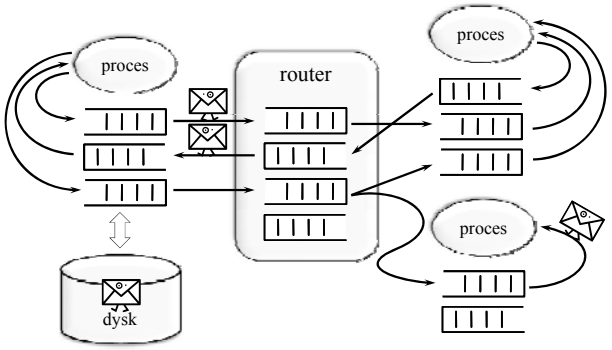
Paradygmat publish/subscribe

- Udostępnienie mechanizmu komunikacji opartego na identyfikacji wiadomości, a nie adresu nadawcy/odbiorcy
- Komunikujące się strony nie znają się wzajemnie
- Strona publikująca (nadawca) udostępnia treść związaną z określonym tematem (ang. topic)
- Środowisko komunikacyjne (usługa) przekazuje treść udostępnionych wiadomości odbiorcom (subskrybentom), którzy zarejestrowali (zapisali) się na dany temat.

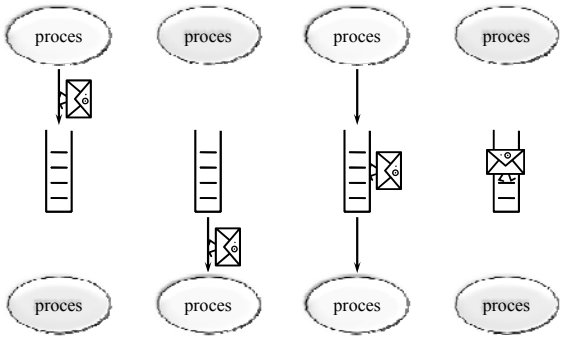
Model systemu kolejkowania komunikatów — podstawowe pojęcia

- Wiadomość (ang. message) — porcja danych (najczęściej z dodatkowymi własnościami) składowana w kolejce
- Kolejka (ang. queue) — miejsce przechowywania wiadomości (komunikatów)
- Proces (ang. process) — element aplikacji, zlecający operacje na wiadomościach w kolejce
- Zarządca zbioru kolejek — moduł na danym węźle, odpowiedzialny za wykonywanie operacji na kolejkach (np. tworzenie, usuwanie, lokalizowanie kolejek, ustawianie atrybutów kolejek itp.)

Model systemu kolejkowania komunikatów — funkcjonowanie



Warianty komunikacji



Przykłady rozwiązań typu MOM

- IBM MQSeries (WebSphere MQ, XMS — Message Service Client)
- Microsoft Message Queuing (MSMQ)
- Java Message Service (JMS)
- Apache ActiveMQ
- Oracle Advanced Queueing
- Sun Java System Message Queue (OpenMQ — wersja open source)
- Usługa IceStorm w systemie ICE

MOM API

rodzaj operacji	IBM MQSeries	MSMQ
tworzenie kolejki		MQCreateQueue
usuwanie kolejki		MQDeleteQueue
otwieranie kolejki	MQopen	MQOpenQueue
zamykanie kolejki	MQclose	MQCloseQueue
wysyłanie wiadomości	MQput	MQSendMessage
odbieranie wiadomości	MQget	MQReceiveMessage

Rozproszona pamięć współdzielona (ang. Distributed Shared Memory)

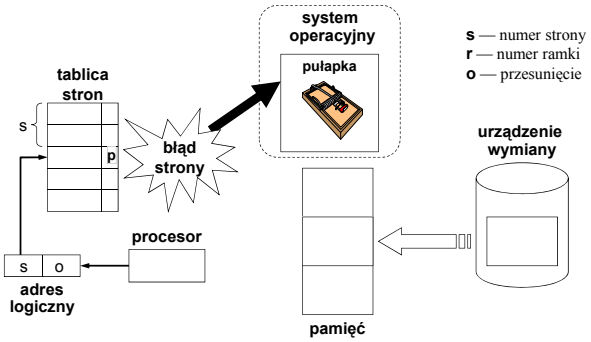
DSM — podstawowe cechy

Cel: dostarczenie wspólnej wirtualnej przestrzeni adresowej, dostępnej (potencjalnie) dla wszystkich węzłów systemu

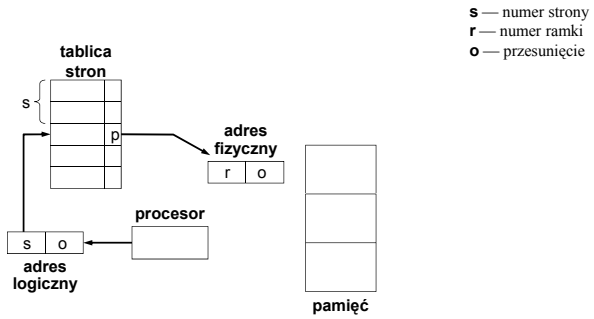
Zalety:

- wygodny dla programisty paradygmat programowania równoległego,
- dostępność wirtualnej przestrzeni adresowej obejmującej pamięci fizyczne wszystkich węzłów,
- możliwość uruchamiania w środowisku rozproszonym programów równoległych zaprojektowanych dla środowiska wieloprocesorowego z pamięcią współdzieloną

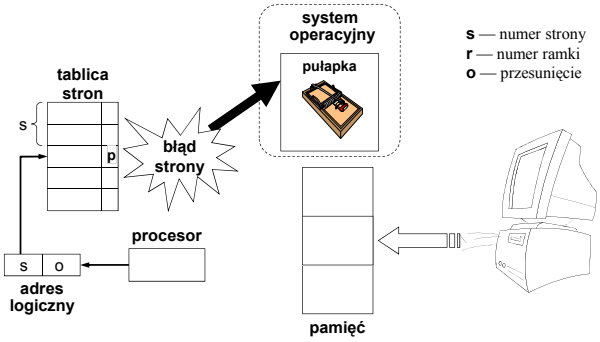
Obsługa błędu strony w systemie pamięci wirtualnej



Obsługa błędu strony w systemie pamięci wirtualnej



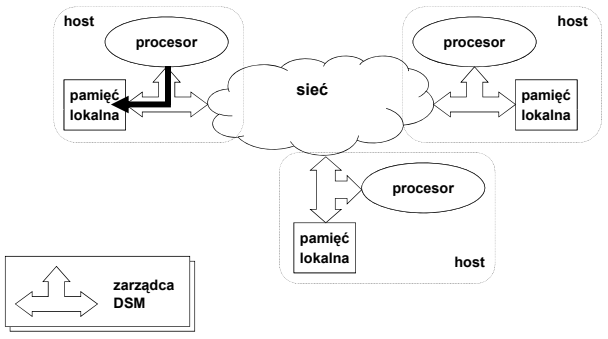
Obsługa błędu strony w systemie wirtualnej pamięci rozproszonej



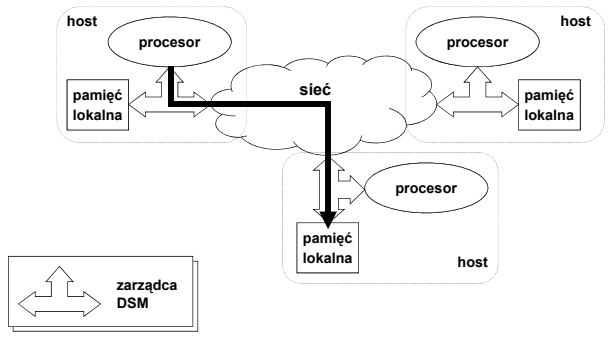
Koncepcje dostępu do danych

- ***Dostęp zdalny*** — każdy dostęp do współdzielonego obiektu, zlokalizowanego fizycznie w pamięci lokalnej innego węzła, odbywa się przez sieć.
- ***Relokacja*** — możliwa jest zmiana fizycznej lokalizacji współdzielonego obiektu, czyli umieszczenie go w pamięci lokalnej węzła, w którym pojawiło się żądanie dostępu.
- ***Replikacja*** — obiekt logiczny może być jednocześnie zlokalizowany fizycznie w pamięci lokalnej wielu węzłów, co umożliwia równoległy dostęp do tego obiektu w wielu węzłach.

Dostęp zdalny



Dostęp zdalny

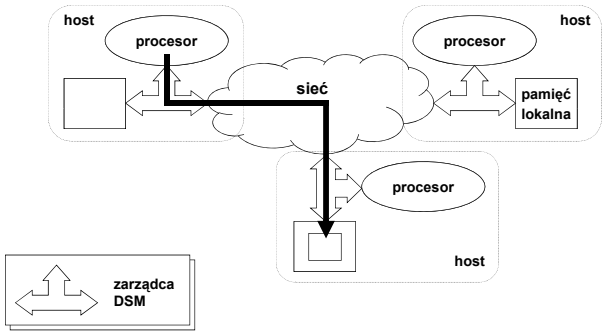


Zdalny dostęp — charakterystyka

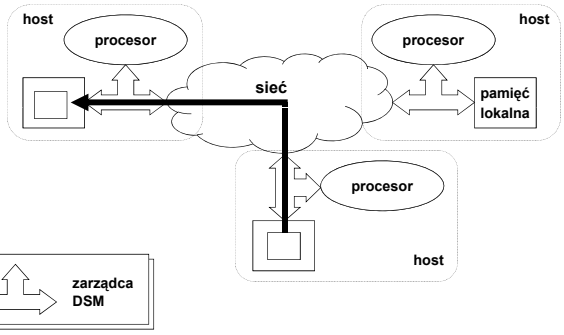
- Stosunkowo prosta realizacja

- Problem efektywności — duży czas dostępu do danych w zdalnych węzłach

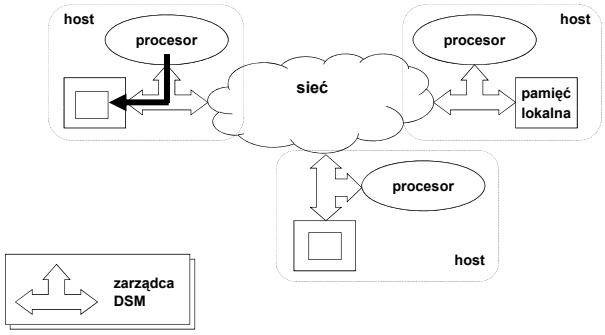
Relokacja



Relokacja



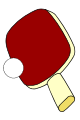
Relokacja



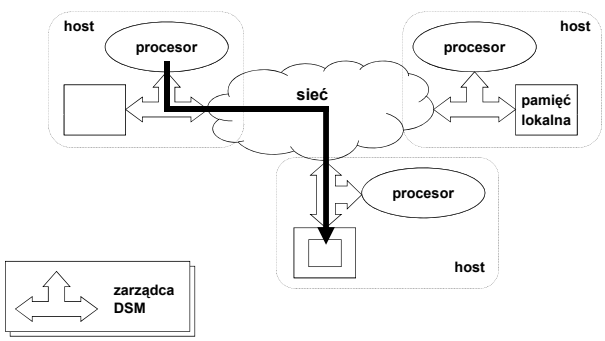
Relokacja — charakterystyka

- Problem lokalizacji
- Problem rozmiaru i struktury jednostki podlegającej relokacji

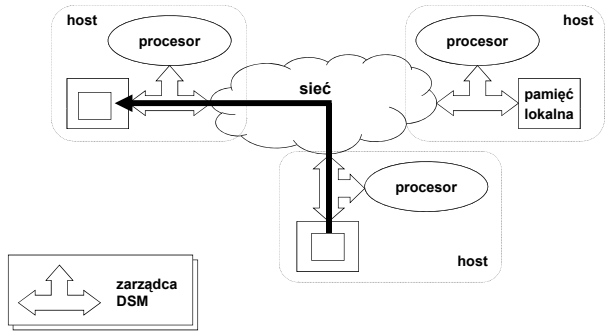
- Problem migotania (ang. trashing, ping-pong effect)



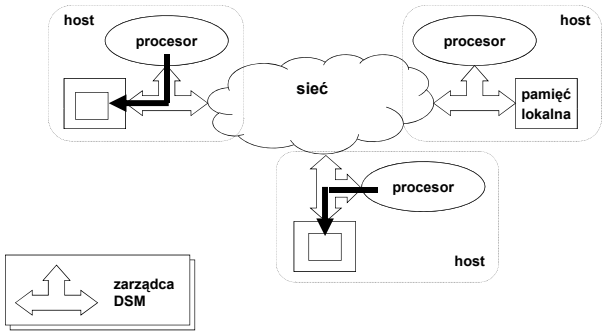
Replikacja



Replikacja

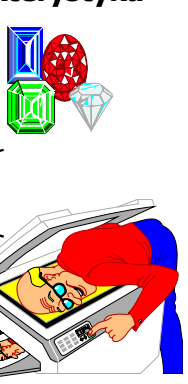


Replikacja



Replikacja — charakterystyka

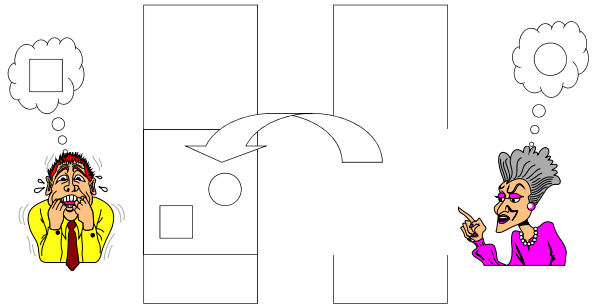
- Problem lokalizacji
- Problem rozmiaru i struktury jednostki podlegającej replikacji
- Problem migotania (ang. trashing, ping-pong effect)
- Problem spójności kopii (replik)



Struktura jednostki podlegającej replikacji lub relokacji

- ***Strona*** — fizyczne połączenie kilku odrębnych obiektów logicznych w jedną jednostkę udostępnianą jako całość przez DSM (problem fałszywego współdzielenia).
- ***Pojedyncza zmienna*** — duży jednostkowy koszt relokacji i utrzymywania spójności.
- ***Obiekt*** — możliwość optymalizacji w strategii utrzymywania spójności w związku ze ściśle określonym sposobem dostępu (poprzez metody).

Fałszywe współdzielenie



Problem spójności replik — protokół koherencji

- Protokół ***unieważniania*** danych (ang. invalidation protocol) — niespójne repliki są usuwane z pamięci lokalnej.
- Protokół ***aktualizacji*** danych (ang. update protocol) — niespójne repliki są aktualizowane.

Problem lokalizacji — system IVY

- ***statyczny scentralizowany*** mechanizm lokalizacji stron
- ***statyczny rozproszony*** mechanizm lokalizacji stron
- ***dynamiczny*** mechanizm lokalizacji stron

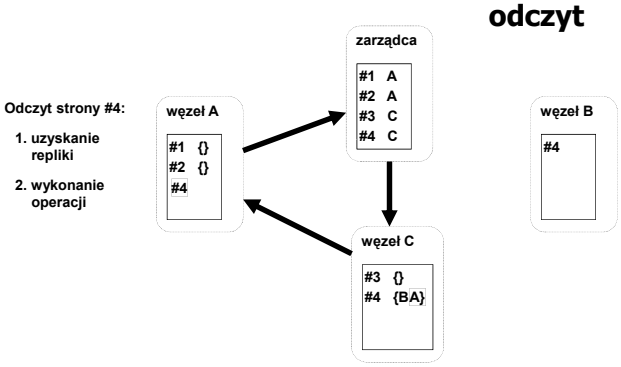
System IVY — podstawowe pojęcia i struktury danych (1)

- właściciel strony — węzeł, na którym była wykonywana ostatnia operacja zapisu danej strony
- zbiór kopii (copyset) — zawiera identyfikatory węzłów posiadających kopię strony (przechowywana przez właściciela strony)

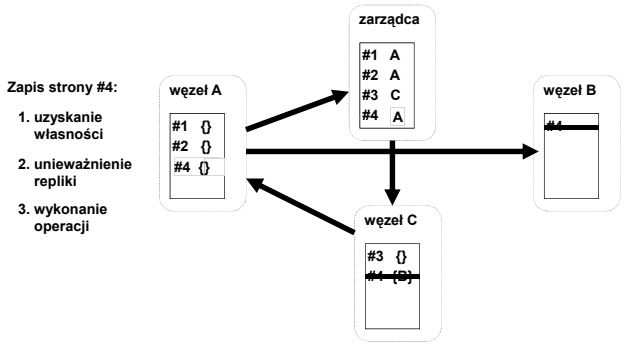
System IVY — podstawowe pojęcia i struktury danych (2)

- zarządca (w podejściu statycznym) — węzeł, który przechowuje dane o właścicielach poszczególnych stron
- tablica właścicieli stron — dla każdej strony zawiera identyfikator jej właściciela (przechowywany przez zarządców)
- prawdopodobny właściciel (w podejściu dynamicznym) — tablica zawierająca dla każdej strony w systemie identyfikator węzła, o którym wiadomo, że był kiedyś (lub jeszcze jest) właścicielem danej strony (przechowywana przez każdy węzeł)

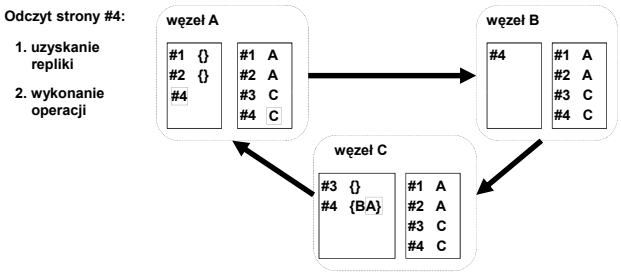
Statyczny scentralizowany mechanizm lokalizacji stron — odczyt



Statyczny scentralizowany mechanizm lokalizacji stron — zapis



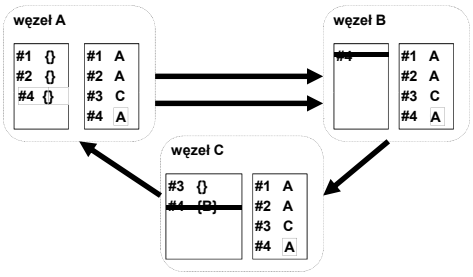
Dynamiczny mechanizm lokalizacji stron — odczyt



Dynamiczny mechanizm lokalizacji stron — zapis

Zapis strony #4:

1. uzyskanie własności
2. unieważnienie repliki
3. wykonanie operacji



Skalowalność

Pojęcie skalowalności

- Skalowalność w sensie rozmiaru instancji problemów — system jest w stanie wykorzystać dodatkowe zasoby na potrzeby świadczenia usług dla zwiększającej się liczby użytkowników lub przy zwiększonej intensywności strumienia zgłoszeń.
- Skalowalność w sensie geograficznym — system jest w stanie świadczyć usługi dla użytkowników znacznie od siebie oddalonych.
- Skalowalność zarządzania (administrowania) — system może być sprawnie zarządzany nawet jeśli swoim działaniem obejmuje wiele różnych domen administracyjnych.

Ograniczenia skalowalności w sensie rozmiaru

- Scentralizowana usługa — usługa udostępniana przez pojedynczy serwer.
  Decentralizacja usługi → zwiększenie narażenia na ataki.
- Scentralizowane dane — dane dostępne na nośniku obsługiwanym przez pojedyncze urządzenie (np. jeden dysk).
  Decentralizacja danych → problem powiązania danych rozproszonych w różnych miejscach.
- Scentralizowane algorytmy — algorytmy realizowane jednowątkowo i wymagające dostępu do całego zbioru przetwarzanych danych.
  Decentralizacja algorytmu → problem synchronizacji wątków i podziału danych

Charakterystyka przetwarzania zdecentralizowanego

- Żadna maszyna nie ma pełnej informacji na temat stanu systemu jako całości.
- Decyzje podejmowane przez poszczególne (kooperujące) procesy podejmowane są na podstawie informacji lokalnych.
- Awaria jednej maszyny nie oznacza konieczności zakończenia przetwarzania w całym systemie.
- Nie ma globalnego zegara synchronizującego przetwarzanie na poszczególnych węzłach.

Ograniczenia skalowalności w sensie geograficznym

- Komunikacja synchroniczna — wydłuża się czas komunikacji i tym samym powstają przestoje wynikające z oczekiwania na potwierdzenie.
- Zawodność komunikacji — zwiększone prawdopodobieństwo utraty komunikatu.
- Komunikacja punkt-punkt — brak efektywnego mechanizmu rozgłaszania.

Ograniczenia skalowalności zarządzania

- Różnice i konflikty w zakresie polityki
  - wykorzystywania zasobów,
  - zarządzania zasobami,
  - bezpieczeństwa.
- W przypadku systemu obejmującego kilka domen administracyjnych należy przygotować osobne mechanizmy do realizacji odpowiedniej polityki w stosunku do użytkowników z danej domeny oraz użytkowników z innych domen.

Techniki poprawy skalowalności

- Ukrywanie opóźnień komunikacyjnych — komunikacja asynchroniczna z ewentualnym wywołaniem zwrotnym w celu przekazania wyniku.
- Migracja przetwarzania — ograniczenie interakcji ze zdalną jednostką poprzez realizację przetwarzania lokalnego.
- Dystrybucja — rozkład obciążenia na różne jednostki współpracujące w ramach systemu rozproszonego.
- Replikacja — jednoczesne wykorzystanie nadmiarowych zasobów w celu powielenia danych lub funkcji systemu tym samym zwiększenia dostępności oraz zrównoleglenia przetwarzania (realizowania tej samej usługi przez wiele serwerów).

Komunikacja asynchroniczna

- Zastosowanie w przypadku przetwarzania wsadowego — wysłanie żądania wykonania pewnego zbioru operacji bez konieczności ingerencji w przetwarzanie.
- Zastosowanie w celu zrównoleglenia przetwarzania — możliwość wysłania żądań do wielu serwerów bez oczekiwania na odpowiedź.
- Ograniczone zastosowania w przypadku aplikacji interaktywnych — interakcja z użytkownikiem wymaga przekazania odpowiedzi i tym samym wymusza komunikację synchroniczną.

Migracja przetwarzania

- Zastosowanie w przypadku przetwarzania interaktywnego — część przetwarzania wstępnego (np. sprawdzenie poprawności wprowadzonych danych w formularzu) może być wykonana lokalnie po stronie klienta i ewentualna konieczność ponownej interakcji z użytkownikiem nie wymaga czasochłonnej komunikacji sieciowej z serwerem.
- Przykład zastosowania w technologiach internetowych:
  - Aplety Java,
  - JavaScript.

Dystrybucja

- Zastosowanie w przypadku przetwarzania zdecentralizowanego — podział komponentów (danych lub kodu) na oddzielne części przetwarzane niezależnie i rozłożenie tych części na różnych węzłach systemu.
- Przykłady zastosowań:
  - usługa DNS (podział danych, czyli dekompozycja dziedziny),
  - usługa NIS+,
  - usługa www.

Cel replikacji

- Zwiększenie dostępności — istnienie wielu replik umożliwia uzyskanie dostępu do usługi nawet w przypadku awarii niektórych serwerów.
- Zwiększenie efektywności — dostępność wielu replik umożliwia rozkład obciążenia żądaniami klientów na wiele jednostek przetwarzających, optymalne zrównoleglenie wymaga jednak zrównoważenia obciążenia.
- Zmniejszenie opóźnień komunikacyjnych — istnienie wielu replik zwiększa szansę na uzyskanie dostępu do serwera zlokalizowanego blisko klienta.
- Przykłady replikacji: NIS, LDAP, DNS, proxy-serwer, system IVY, macierze dyskowe (RAID).

Podejścia do replikacji (1)

- przedmiot replikacji
  - replikacja przetwarzania (procesów, usług) — realizacja tej samej usługi przez kilka procesów w oparciu o wspólny zbiór zasobów
  - replikacja danych (zasobów) — powielanie danych w wielu miejscach (na wielu nośnikach)
- sposób aktualizacji replik
  - transfer stanu — jedna z replik przetwarza zlecenie, a ewentualnie zmieniony w wyniku tego stanu zasobów kopiowany jest do pozostałych replik
  - transfer operacji — zlecenie (zwłaszcza modyfikujące stan zasobu) przekazywane jest do wszystkich replik i każda z nich je przetwarza

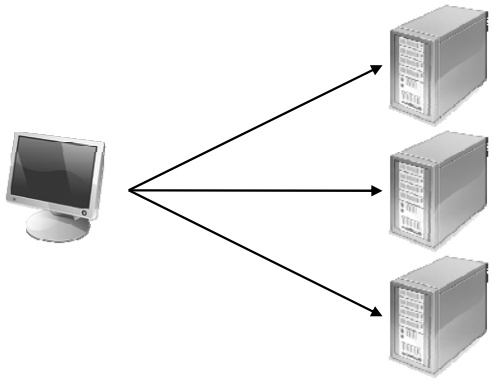
Podejścia do replikacji (2)

- sposób udostępniania replik
  - jedna kopia podstawowa — użytkownicy mają dostęp do jednej wybranej repliki, pozostałe repliki stanowią kopie zapasowe
  - wiele kopii podstawowych — kilka replik (w szczególności wszystkie) mogą przyjmować zlecenia bezpośrednio od użytkowników
- zarządzanie spójnością replik
  - replikacja pesymistyczna — różnica w stanie replik jest ukrywana przed użytkownikiem
  - replikacja optymistyczna — różnica w stanie replik może się ujawnić, ale zakłada się optymistycznie, że ten fakt nie wpłynie na poprawność funkcjonowania systemu lub zostanie nie zauważony

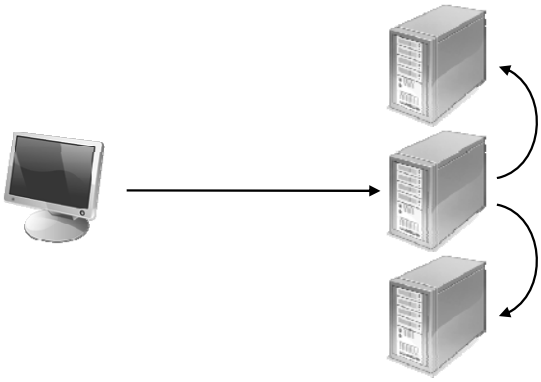
Techniki replikacji

	transfer operacji	transfer stanu
jedna kopia podstawowa	półaktywna	pasywna
wiele kopii podstawowych	aktywna	półpasywna

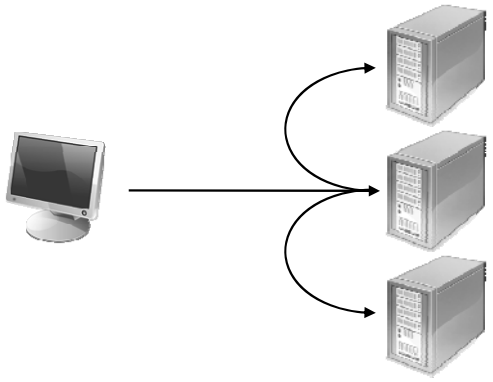
Replikacja aktywna



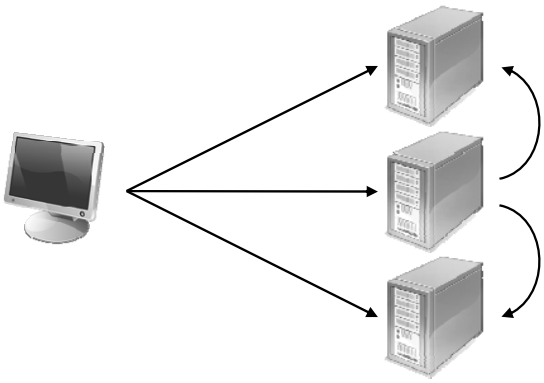
Replikacja pasywna



Replikacja półaktywna



Replikacja półpasywna



Protokół aktualizacji

- Wypychanie aktualizacji — przekazywanie danych do replik inicjowane jest przez serwer wprowadzający modyfikację zaraz po zmianie stanu
- Wyciąganie aktualizacji — zamówienie uaktualnień inicjowane jest przez serwer repliki i przekazywane do serwera (serwerów), który wprowadził zmiany
- Powiadamianie o aktualizacjach — przekazywanie informacja o zasobach (danych), których stan się zmienił, inicjowane przez serwer wprowadzający modyfikację zaraz po zmianie stanu — może skutkować unieważnieniem danych w replikach lub zainicjować wyciąganie aktualizacji

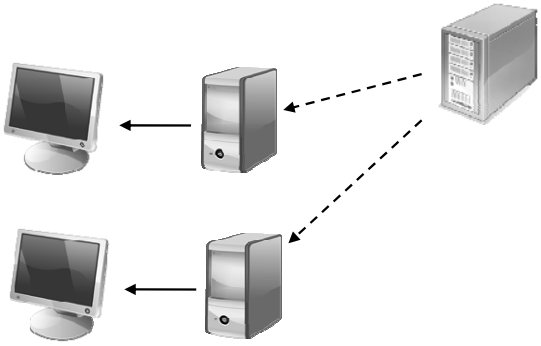
Rodzaje replik

- repliki trwałe — kopie tworzone statycznie przez twórcę usługi (właściciela/dystrybutora replikowanych danych) i istniejące przez cały czas funkcjonowania usługi
- repliki inicjowane przez serwer — tworzone dynamicznie z inicjatywy właściciela/dystrybutora (replikowanych) danych w celu poprawy efektywności
- repliki inicjowane przez klienta — kopie podręczne tworzone lokalnie na potrzeby określonego klienta lub niewielkiej grupy klientów → buforowanie podręczne

Buforowanie podręczne (ang. caching)

- Buforowanie podręczne jest formą replikacji polegającą na tworzeniu replik możliwie blisko lokalizacji klienta, który ma z niej korzystać.
- Różnica pomiędzy repliką a kopią podręczną:
  - replika tworzona jest przez właściciela zasobu na użytek wielu klientów,
  - kopia podręczna tworzona jest przez klienta na jego prywatne potrzeby.
- Przykład buforowania podręcznego: przeglądarka www, klient poczty elektronicznej.

Buforowanie podręczne

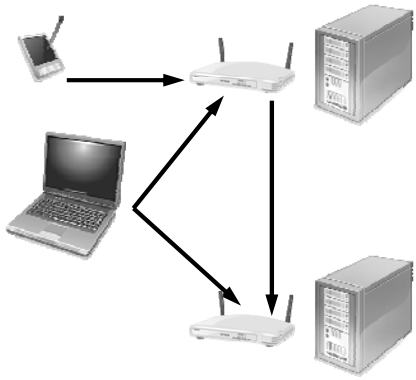


Problem spójności replik

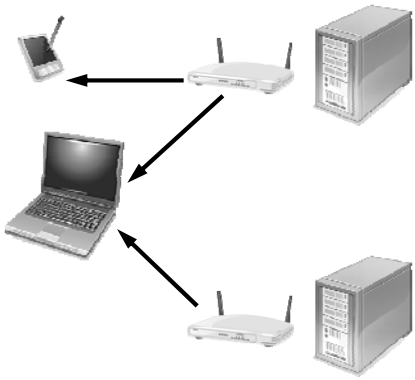
- W przypadku replik zarządzanie spójnością może być realizowane po stronie serwerów, które są świadome istnienia wielu kopii.
- W przypadku kopii podręcznej odpowiedzialność za spójność danych spoczywa na klientach.
- Utrzymanie ścisłej spójności replik przy dużej liczbie modyfikacji może być dość kosztowne, dlatego stosowana jest replikacja optymistyczna.

Gwarancje dla sesji klienta

Operacja aktualizacji — zapis



Operacja odczytu



Notacja

$w(x)v$ —	operacja zapisu obietu x
$r(x_i)v$ —	operacja odczytu repliki na serwerze S_i
$o1 \rightarrow_{Cj} o2$	— kolejność zgłaszania operacji przez klienta C_j
$o1 \rightarrow_{Si} o2$	— kolejność wykonania operacji przez serwer S_i

Gwarancje dla sesji klienta (modele spójności zorientowane na klienta)

- odczyt własnych zapisów (ang. read your writes)
- monotoniczność zapisów (monotonic writes, FIFO of writes)
- następstwo zapisów po odczytach (writes follow reads, reads before writes, session causality)
- monotoniczność odczytów (monotonic reads, FIFO of reads)

Założenie ogólne

Operacje zapisu tej samej zmiennej wykonywane są w tej samej kolejności przez wszystkie serwery (koherencja)

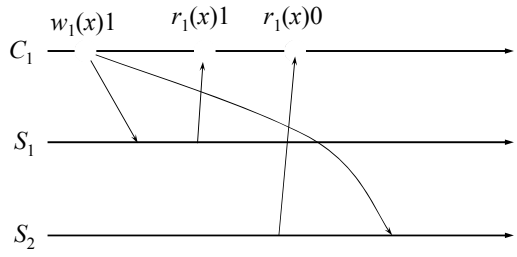
$$\forall_x \left(\forall_{S_i} w(x)v \rightarrow_{S_i} w(x)u \vee \forall_{S_i} w(x)u \rightarrow_{S_i} w(x)v \right)$$

Odczyt własnych zapisów

Skutek operacji zapisu będzie widoczny w następnych operacjach odczytu zgłoszonych przez tego samego klienta

$$\forall_{C_i} \forall_{S_j} \left(w(x)v \rightarrow_{C_i} r(y_j)u \Rightarrow w(x)v \rightarrow_{S_j} r(y)u \right)$$

Odczyt własnych zapisów — przykład

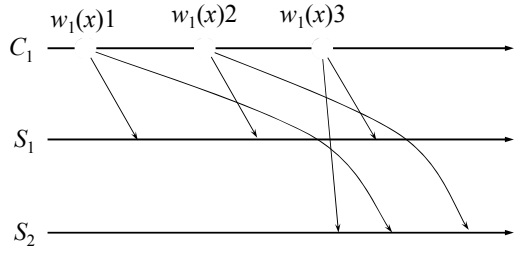


Monotoniczność zapisów

Operacje zapisu danego procesu wykonywane są na każdym serwerze w kolejności ich zgłoszenia

$$\exists_{C_i} w(x)v \rightarrow_{C_i} w(y)u \Rightarrow \forall_{S_j} w(x)v \rightarrow_{S_j} w(y)u$$

Monotoniczność zapisów — przykład

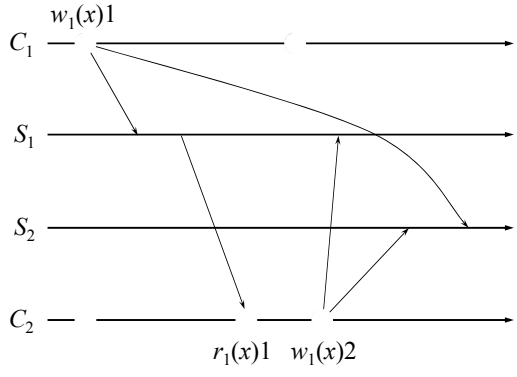


Następstwo zapisów po odczytach

Operacja zapisu zgłoszona przez klienta po odczytaniu określonej wartości innej zmiennej może być wykonana na serwerze, na którym została już zapisana wartość odczytana przez danego klienta.

$$\exists_{C_i} r(x)v \rightarrow_{C_i} w(y)u \Rightarrow \forall_{S_j} w(x)v \rightarrow_{S_j} w(y)u$$

Następstwo zapisów po odczytach — przykład

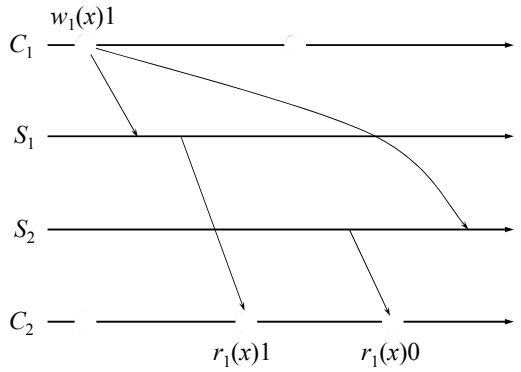


Monotoniczność odczytów

Operacja odczytu może być wykonana tylko na serwerze, na którym wykonane były zapisy wcześniej odczytanych wartości

$$\forall_{C_i}\forall_{S_j}\left(r(x)v \rightarrow_{C_i} r(y_j)u \Rightarrow w(x)v \rightarrow_{S_j} r(y)u\right)$$

Monotoniczność odczytów — przykład



Implementacja gwarancji sesji

- Implementacja naiwna — identyfikowanie operacji i tworzenie zbiorów identyfikatorów operacji zgłoszonych do systemu, oraz operacji mających wpływ na bieżący stan przetwarzania.
- Implementacja wektorowa — utrzymywanie liczników operacji zgłoszonych oraz liczników operacji wykonanych w systemie.

Implementacja naiwna gwarancji sesji — struktury danych

- Po otrzymaniu żądania wykonania operacji zapisu serwer nadaje identyfikator operacji i zwraca go klientowi.
- Klient C_i utrzymuje dwa zbiory operacji:
  - R_i — zbiór operacji zapisu, których wykonania sam zażądał — aktualizowany po otrzymaniu identyfikatora od serwera,
  - W_i — suma zbiorów operacji, które wykonały poszczególne serwery do momentu ostatniej operacji odczytu przez danego klienta — aktualizowany po otrzymaniu wyniku operacji odczytu.
- Serwer S_j utrzymuje zbiór WS_j operacji zapisu, które wykonał bezpośrednio (jako wynik żądania klienta) lub pośrednio (w wyniku aktualizacji replik).

Implementacja naiwna gwarancji sesji — odczyt własnych zapisów

- Klient C_i wysyła do serwera S_j zbiór R_i
- Serwer może wykonać operację, gdy $R_i \subseteq SW_j$
- Po wykonaniu operacji serwer zwraca klientowi zbiór SW_j
- Po odebraniu wyniku klient aktualizuje zbiór W_i
  $W_i \leftarrow W_i \cup WS_j$

Implementacja naiwna gwarancji sesji — monotoniczność zapisów

- Klient C_i wysyła do serwera S_j zbiór R_i
- Serwer może wykonać operację, gdy $R_i \subseteq SW_j$
- Po wykonaniu operacji serwer zwraca klientowi identyfikator operacji *opid* oraz aktualizuje zbiór SW_j
  $SW_j \leftarrow SW_j \cup \{opid\}$
- Po odebraniu wyniku klient aktualizuje zbiór R_i
  $R_i \leftarrow R_i \cup \{opid\}$

Implementacja naiwna gwarancji sesji — monotoniczność odczytów

- Klient C_i wysyła do serwera S_j zbiór W_i
- Serwer może wykonać operację, gdy $W_i \subseteq SW_j$
- Po wykonaniu operacji serwer zwraca klientowi zbiór SW_j
- Po odebraniu wyniku klient aktualizuje zbiór W_i
  $W_i \leftarrow W_i \cup WS_j$

Implementacja naiwna gwarancji sesji — następstwo zapisów po odczytach

- Klient C_i wysyła do serwera S_j zbiór W_i
- Serwer może wykonać operację, gdy $W_i \subseteq SW_j$
- Po wykonaniu operacji serwer zwraca klientowi identyfikator operacji *opid* oraz aktualizuje zbiór SW_j
  $SW_j \leftarrow SW_j \cup \{opid\}$
- Po odebraniu wyniku klient aktualizuje zbiór R_i
  $R_i \leftarrow R_i \cup \{opid\}$