IBM DCE/DFS

Krzysztof Onak
konak@mimuw.edu.pl

18 grudnia 2002

1 IBM DCE

1.1 Ogólne spojrzenie

IBM Distributed Computing Environment (DCE) (Środowisko Obliczeń Rozproszonych) dostarcza usługi i narzędzia, które wspomagają tworzenie, używanie i zarządzanie aplikacjami rozproszonymi w niejednorodnym środowisku obliczeniowym. DCE składa się z komponentów, które dzielą na dwie podstawowe kategorie:

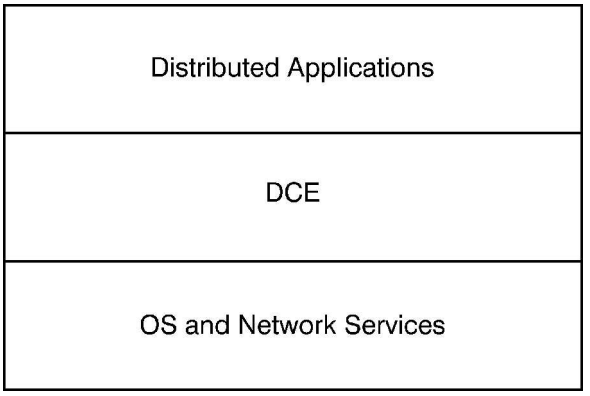
- narzędzia — asystują przy tworzeniu aplikacji;
- usługi — które dostarczają wsparcie potrzebne w środowisku rozproszonym (analogia wsparcia, które daje system operacyjny w scentralizowanym systemie).

DCE dostarcza dostarcza zintegrowany zestaw usług i wspomaga współdzielenie danych.

DCE jest oparty na trzech modelach obliczeń rozproszonych. Są to:

- model klient/serwer;
- model zdalnego wywoływania procedur (ang. remote procedure call);
- model współdzielenia danych.

IBM dostarcza implementacje DCE dla systemów operacyjnych AIX i Solaris.



Rysunek 1: Schemat warstw

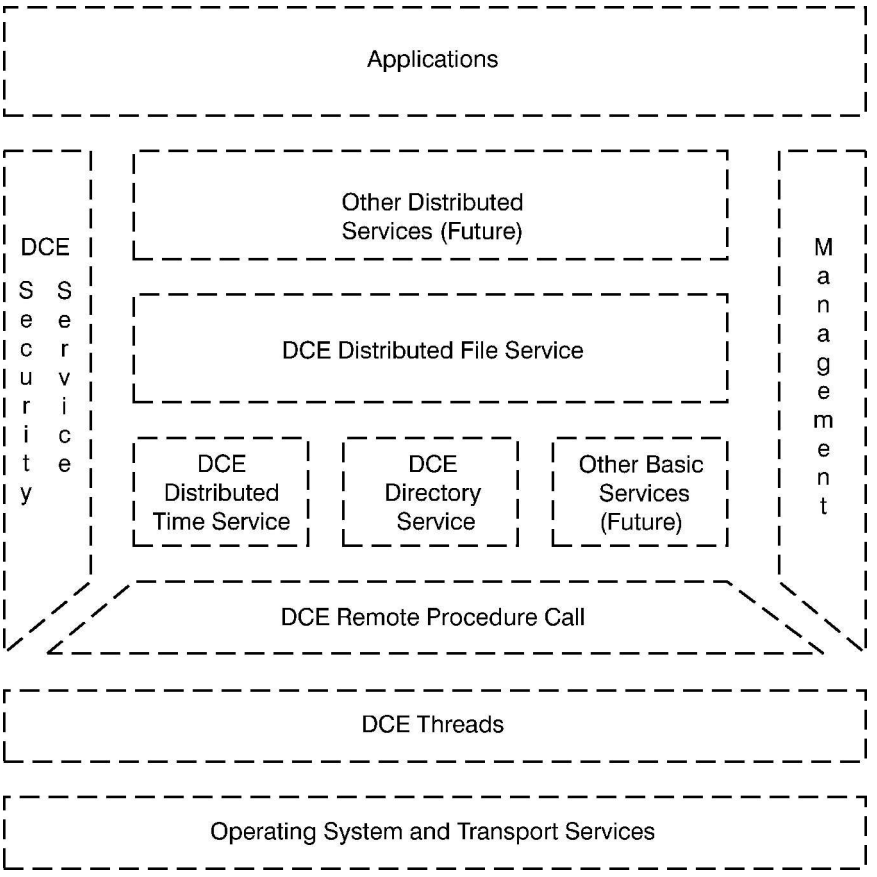
1.2 Architektura DCE

IBM DCE stanowi warstwę pomiędzy systemem operacyjnym i siecią z jednej strony a rozproszonymi aplikacjami z drugiej strony. DCE dostarcza usługi, które pozwalają rozproszonym aplikacjom na interakcję z potencjalnie różnymi komputerami, systemami operacyjnymi i sieciami, jakgdyby stanowiły jeden system. Rysunek 1 obrazuje DCE w relacji do systemów operacyjnych, oprogramowania komunikacji sieciowej i aplikacji.

Wiele różnych komponentów współpracuje razem, aby zaimplementować warstwę DCE. Wiele z nich dostarcza w środowisku rozproszonym funkcje, które w zwykłym scentralizowanym środowisku dostarcza system operacyjny. Na rysunku 2 przedstawiony jest zgrubny schemat architektury DCE. W dalszej części dokumentu skupimy się na komponencie DCE Distributed File Service.

2 IBM DFS

DFS (ang. Distributed File System — rozproszony system plików) to aplikacja typu klient/server, która udostępnia globalną przestrzeń plików i katalogów w środowisku DCE. Przestrzeń ta jest niezależna od fizycznego położenia plików. Pliki mogą być przechowywane na wielu różnych komputerach, ale być dostępne dla wszystkich użytkowników w sieci. Niezależnie od tego użytkownik otrzymuje pojedynczą przestrzeń plików.



Rysunek 2: Architektura DCE

2.1 Przegląd dostarczanych przez IBM implementacji

IBM dostarcza implementacje serwerów i klientów DFS dla systemów operacyjnych Solaris i AIX. Dostępne są także usługi klienckie dla komputerów z Windows NT.

2.2 Serwery DFS

Na serwerach DFS działają procesy zajmujące się takimi sprawami, jak udostępnianie danych oraz monitorowanie i kontrolowanie innych procesów. Serwery dzielą się na kategorie w zależności od tego, jakie procesy na nich działają. Krótki opis kategorii:

- **Serwer Plików** (ang. File Server machine) to komputer, na którym są uruchomione procesy odpowiedzialne za przechowywanie i eksportowanie danych. Są to na przykład Serwer Zestawu Plików (ang. Fileset Server), który udostępnia interfejs do poleceń systemu plików DFS i komponenty używane do manipulacji plikami, i Eksporter Plików (ang. File Exporter), który działa w zmodyfikowanym jądrze systemu i udostępnia pliki systemu DFS globalnej przestrzeni nazw DCE.

- **Maszyna Kontrolująca System** (ang. System Control machine) zajmuje się tym, aby na wszystkich komputerach były dostępne identyczne wersje plików konfiguracyjnych.

- **Maszyna Dystrybucji Binariów** (ang. Binary Distribution machine) rozprowadza w sieci binaria systemu.

- **Maszyna Bazy Danych Zestawu Plików** (ang. Fileset Database machine) przechowuje główną wersję bądź kopię Bazy Danych Lokacji Zestawów Plików (ang. FLDB = Fileset Location Database), w której znajduje się informacja o położeniu plików systemowych i użytkownika.

- **Maszyna Bazy Danych Kopii Zapasowych** (ang. Backup Database machine) przechowuje główną wersję i kopie Bazy Danych Kopii Zapasowych, w której znajdują się informacje używane do przywracania plików systemowych i użytkownika.

Jedna maszyna może należeć do kilku kategorii i pełnić wiele funkcji.

2.3 Komputery klienckie DFS

Komputery klienckie DFS udostępniają moc obliczeniową, dostęp do plików w systemie DFS i inne narzędzia ogólnego przeznaczenia. Można skonfigurować system tak, aby serwery służyły także jako komputery klienckie.

Jądro systemu na maszynach klienckich jest zmodyfikowane i zarządza połączeniami z Eksporterem Plików i procesami serwerowymi. Zbiór modyfikacji jądra nazywa się Zarządca Pamięci Podręcznej (ang. Cache Manager). Główne zadanie Zarządcy Pamięci Podręcznej to tłumaczenie żądań aplikacji na maszynie klienckiej skierowanych do systemu plików na zdalne wywołania procedur (ang. remote procedure calls) skierowane do procesów Eksporterów Plików uruchomionych na Serwerach Plików.

Kiedy Zarządca otrzymuje żądane dane od Eksportera Plików zachowuje je w pamięci podręcznej przed przekazaniem do aplikacji, która ich żądała. Na dodatek DFS zapewnia, że Zarządca ma zawsze dostęp do najbardziej aktualnej kopii danych. Jeśli centralna kopia pliku przechowującego dane zmienia się, Zarządca pobiera nową wersję pliku przy następnym odwołaniu do tego pliku. Użytkownik nie musi zlecać Zarządcy utrzymywania aktualnej kopii — działanie Zarządcy jest automatyczne i całkowicie przeźroczyste dla użytkownika.

2.4 Zarządzanie dostępem do danych

Aby zsynchronizować dostęp do danych, Eksporter Plików na każdym Serwerze Plików rozdziela żetony (ang. tokens) klientom, którzy korzystają z plików na serwerze. Eksporter Plików używa żetonów do zarządzania dostępem do danych i metadanych. Żetony gwarantują, że każdy klient pracuje z najbardziej aktualną wersją danych oraz że klienci nie korzystają z danych w sposób mogący powodować konflikty. Żetony są całkowicie przeźroczyste zarówno dla użytkowników jak i dla administratorów.

Kiedy klient, taki jak Zarządca Pamięci Podręcznej, potrzebuje pobrać lub zmienić plik lub katalog zarządzany przez Eksportera Plików, najpierw prosi o odpowiedni żeton na dane z Eksportera Plików. Odpowiedź Eksportera na prośbę klienta zależy od danych, które klient żąda, operacji, którą klient chce przeprowadzić na danych, i od tego, czy inni klienci posiadają aktualnie żetony na dane.

Jeśli inni klienci nie posiadają żetonów na te dane, Eksporter Plików może wydać klientowi odpowiednie żetony. Jeśli istnieją wydane żetony na dane, Eksporter może spełnić prośbę (o ile nie zachodzą konflikty pomiędzy żądaniem a wydanymi żetonami), odzyskać wydane żetony i spełnić prośbę lub odłożyć prośbę do czasu, gdy będzie mógł ją spełnić. W pewnych przypad-

kach Eksporter Plików po prostu odmawia spełnienia prośby. Jeśli Eksporter wydał klientowi potrzebny żeton, to klient może korzystać z danych w sposób, którego żądał.

2.5 Domeny administracyjne DFS

W DCE podstawową jednostką operacyjną jest komórka (ang. cell). Komórka składa się z jednego do kilku tysięcy systemów dzielących administracyjnie niezależne instalacje serwerów i komputerów klienckich, zunifikowane środowisko nazw i wspólny serwer autoryzacji i bazę danych. Zarówno wiele komórek może istnieć w jednym miejscu, jak i odległe maszyny mogą należeć do tej samej komórki, jednakże jeden komputer może należeć do jednej komórki.

Użytkownik może mieć dostęp do wielu komórek. Jednakże uniwersalny identyfikator użytkownika pojawia się w rejestrze tylko jednej komórki. Ta komórka jest komórką lokalną (lub komórką domową) użytkownika. Wszystkie pozostałe komórki są obcymi komórkami zarówno z perspektywy użytkownika, jak i maszyn w komórce domowej użytkownika.

Podczas logowania na komputerze użytkownik autoryzuje się w komórce, do której komputer należy. Jeśli maszyna należy do komórki domowej użytkownika, identyfikator użytkownika pojawia się w rejestrze komórki. Jeśli komputer jest w obcej komórce, to identyfikator użytkownika nie pojawia się w rejestrze komórki. Musi natomiast istnieć wzajemne zaufanie pomiędzy obydwoma komórkami, aby nastąpiła autoryzacja użytkownika. Administrator systemu, który konfiguruje komórkę, określa, czy bierze ona udział w usłudze globalnego adresowania. Jeśli komórka użytkownika bierze udział w usłudze globalnego adresowania, to może on pozwolić użytkownikom innej komórki, która także bierze udział w usłudze globalnego adresowania, na dostęp do swoich danych, o ile istnieje wzajemne zaufanie pomiędzy komórkami.

DFS rozszerza dalej koncepcję komórek DCE dostarczając domeny administracyjne. Domena administracyjna jest kolekcją połączonych serwerów z tej samej komórki skonfigurowaną do administrowania jako pojedyncza jednostka. Komórka może zawierać wiele maszyn. Domeny administracyjne upraszczają administrację systemem DFS w pojedynczej komórce DCE, organizując podzbiór maszyn komórki w mniejsze jednostki administracyjne. W dodatku do uproszczenia administracji DFS w komórce, domeny wnoszą wysoki poziom elastyczności do administracji DFS w ogólności.

Komórka może zawierać jedną lub więcej domen administracyjnych. Domena administracyjna, tak jak komórka, może zawierać wiele serwerów, które wykonują wspomniane wcześniej zadania. Komputer może może należeć do

wielu domen, ale wszystkie maszyny w domenie muszą należeć do tej samej komórki. Na przykład wszystkie domeny w komórce mogą używać tej samej Maszyny Dystrybucji Binariów, ale maszyna musi być w tej samej komórce, w której są wszystkie maszyny wszystkich domen. Domeny administracyjne są niewidoczne z punktu widzenia użytkownika niebędącego administratorem.

2.6 Listy administracyjne i grupy DFS

Listy administracyjne są plikami, które są używane do wyspecyfikowania, którzy użytkownicy mogą wydawać polecenia, które wpływają na określone procesy i dane w domenie administracyjnej. Bycie członkiem listy administracyjnej oznacza posiadanie pozwoleń potrzebnych do wydawania poleceń na skojarzonym procesie serwerowym. Indywidualni użytkownicy, grupy i serwery mogą być umieszczani na listach administracyjnych. Użytkownik, który jest uczestnikiem grupy znajdującej się na liście administracyjnej ma wszystkie prawa skojarzone z listą. Członkowie grup mają wszystkie prawa wszystkich list, które zawierają grupę. Administracja jest uproszczona przez zmianę członkostwa w grupach i praw nadanych listom zamiast nadawania praw bezpośrednio indywidualnym użytkownikom.

2.7 Lokalny system plików DCE

Lokalny system plików DCE (ang. DCE Local File System = DCE LFS) jest wydajnym systemem plików opartym na logach. DCE LFS korzysta z agregatów. Fizycznie agregat jest równoważny standardowej partycji uniksowej dysku, ale zawiera także specjalizowane metadane o strukturze i lokacji informacji na agregacie. DCE LFS zarządza zbiorem logów wszystkich modyfikacji metadanych, takich jak utworzenie pliku i modyfikacje. Log jest zupełnie niewidoczny dla użytkownika i nie wymaga specjalnej administracji. W przypadku nienormalnego zamknięcia systemu, DCE LFS odtwarza z logów informację o metadanych i używa jej do przywrócenia agregatu do spójnej postaci.

Aby bardziej zapewnić spójność systemu plików po nienormalnym zamknięciu DFS dostarcza program DFS Salvager. W przypadku, gdy odtworzenie logów nie pozwala na odzyskanie spójności systemu, korzysta się właśnie z tego programu. Sposób działania logów DFS i DFS Salvager można porównać do działania rozszerzonego programu fsck, z tym, że fsck używa się także do sprawdzania spójności systemu plików, a Salvager jest uruchamiany tylko do naprawiania systemu plików.

Agregaty DCE LFS wspierają także użycie zestawów plików (ang. filesets). Zestaw plików DCE LFS to hierarchicznie pogrupowane pliki zarzą-

dzane jako pojedyncza jednostka. Zestawy plików mogą się różnić w rozmiarze, ale niemal zawsze są mniejsze niż partycja dysku. Wiele zestawów plików może być przechowywanych na pojedynczym agregacie, co pozwala na elastyczne użycie pamięci dyskowych. Partycja niebędąca LFS (np. partycja uniksowa) może być wyeksportowana do przestrzeni nazw dla użytku jako agregat z DFS, ale może zawierać tylko jeden zestaw plików, niezależnie od ilości danych rzeczywiście przechowywanych w zestawie.

Unikalna struktura metadanych agregatów DCE LFS pozwala na dodatkowe operacje, które nie występują na zwykłych partycjach. Z pomocą DCE LFS potencjalnie małe rozmiary zestawów plików pozwalają na łatwe zarządzanie nimi w celu uzyskania maksymalnej efektywności systemu. Administrator systemu może przenosić zestawy plików z jednego agregatu do innego lub z jednej maszyny na drugą, aby zmniejszyć ilość przesyłanych danych pomiędzy maszynami. Jeśli cała zawartość katalogu domowego użytkownika jest przechowywana w jednym zestawie danych, to cały katalog zostaje przeniesiony, kiedy cały zestaw danych jest przenoszony.

Każdy zestaw plików odpowiada logicznie drzewu katalogów w systemie plików. Każdy zestaw danych zarządza, na pojedynczym agregacie, wszystkimi danymi, które tworzą pliki w drzewie katalogu.

Miejsce, w którym zestaw plików jest podczepiany do globalnej przestrzeni plików, nazywa się punktem montowania (ang. mount point). Punkt montowania wygląda i zachowuje się jak główny katalog zestawu plików. Ta odpowiedniość pomiędzy katalogiem i zestawem plików także upraszcza proces lokacji pliku. Punkt montowania identyfikuje zestaw plików przez nazwę, zatem DFS może automatycznie odnaleźć zestaw plików, nawet jeśli ten jest przenoszony pomiędzy agregatami i/lub maszynami.

Każdy zestaw plików ma przypisaną kwotę (ang. quota), która określa maksymalną ilość pamięci dyskowej, którą mogą zawierać dane zestawu pliku.

DCE LFS pozwala także używanie DCE ACL (Acccess Control Lists), które pozwalają na ustawianie praw na katalogi i pliki w systemie plików LFS. DCE ACL rozszerzają standardowy uniksowy model praw dostępu do pliku i pozwalają na dokładniejszą specyfikację praw dostępu do katalogów i plików. Listy ACL i listy administracyjne wspólnie ograniczają dostęp do operacji zarządzania w komórce lub domenie.

2.8 Kopiowanie zestawów plików

DCE LFS pozwala na kopiowanie zestawów plików DCE LFS. Kopiowanie polega na umieszczeniu kopii zestawów danych na wielu wielu serwerach. Niedostępność pojedynczego serwera przechowującego skopiowany zestaw zwykle nie oznacza przerwania pracy korzystającej z tego zestawu, ponieważ kopie

zestawu są wciąż dostępne na innych maszynach. Np. kopiowanie często używanych plików konfiguracyjnych i plików z binariami znacząco zmniejsza szansę na ich niedostępność jako wynik awarii jednego z serwerów. Kopiowanie przeciwdziała również przeciążeniu serwera udostępniającego często używany zestaw plików. Z kopiowania można korzystać tylko w przypadku zestawów plików DCE LFS.

Są dostępne dwa rodzaje kopiowanie. Dla każdego zestawu plików administrator może wybrać, który rodzaj zastosować. Oto one:

- **Kopiowanie Edycji** (ang. Release Replication) powoduje wydawanie polecenia zaktualizowania przeznaczonej tylko do odczytu kopii, gdy musi ona odzwierciedlać aktualny stan zestawu plików do odczytu i zapisu. Ten typ kopiowania jest użyteczny, gdy rzadko następują zmiany lub gdy zmiany w plikach muszą być wszędzie natychmiast widoczne.

- **Kopiowanie Planowe** (ang. Scheduled Replication) powoduje aktualizowanie co pewien czas kopii nowymi wersjami zestawów plików. Ten typ kopiowania jest przydatny, gdy preferowana jest automatyzacja procesu i nie wszędzie muszą być natychmiast dostępne ostatnie edycje pliku.

Oba typy kopiowanie powodują ostatecznie, że źródłowe zestawy plików są kopiowane na różne serwery.

2.9 System Kopii Zapasowych DFS

DFS dostarcza dwie metody zarządzania kopiami zapasowymi: System Kopii Zapasowych (ang. DFS Backup System) i zestawy plików kopii zapasowych. Korzystając z Systemu Kopii Zapasowych można skopiować dane z zestawów plików na taśmę i odtworzyć w przypadku utraty danych. Informacja o kopiach zapasowych i taśmach jest przechowywana w Bazie Danych Kopii Zapasowych. Sama baza może zostać skopiowana na taśmę i odzyskana w przypadku uszkodzenia. Wspierane jest tworzenie kopii zapasowych zestawów plików.

Można przeprowadzać zarówno pełne jaki i przyrostowe tworzenie kopii zapasowych. Pełne kopiowanie kopiuje wszystkie dane na taśmę, a przyrostowe zapisuje jedynie zmiany od czasu ostatniego tworzenia kopii zapasowej.

Odtwarzanie danych z taśmy przebiega w podobny sposób. Operacja pełnego odtwarzania odtwarza dane według ostatniej pełnej kopii zapasowej uwzględniając późniejsze kopie przyrostowe. Można odtwarzać także pojedyncze pliki, wszystkie zestawy plików znajdujące się na określonym agregacie lub wszystkie pliki, które spełniają wyspecyfikowane kryteria (np. zestawy

plików, które znajdują się na określonym Serwerze Plików lub posiadają określony prefiks).

System Kopii Zapasowych pozwala także na używanie zautomatyzowanych napędów kopii zapasowych. Odpowiednio konfigurując system możemy pozwolić Systemowi Kopii Zapasowych na automatyczne wykonywanie kopii zapasowych.

2.10 Bazy danych DFS

DFS przechowuje informację o zestawach plików w dwóch administracyjnych bazach danych. Baza Danych Lokacji Zestawów Danych (ang. FLDB = Fileset Location Database) przechowuje informację o położeniu zestawów plików, a Baza Danych Kopii Zapasowych zapisuje informacje o kopiach zapasowych i taśmach. Aby zwiększyć niezawodność systemu i dostępność te dwie bazy danych mogą być skopiowane na wiele serwerów. Jeśli jedna z maszyn jest niedostępna, to wciąż baza danych jest dostępna z innych maszyn. W przypadku, gdy jeden z serwerów FLDB przestaje być dostępny, Zarządca Pamięci Podręcznej próbuje łączyć się z kolejnym preferowanym adresem maszyny FLDB. Domyślnie preferencje są dobierane tak, aby robić sensowne decyzje kolejności próbowanych serwerów. Na przykład Zarządca próbuje się najpierw łączyć z maszynami w tej samej podsieci zanim skontaktuje się z maszynami w innych podsieciach.

Aby synchronizować informację w bazach danych, DFS używa biblioteki narzędzi zwanej *Ubik*. Ubik jest mechanizmem synchronizacyjnym, który dystrybuuje zmiany na zestawach plików i kopiach informacji do wszystkich baz danych. Administratorzy muszą wiedzieć, które maszyny przechowują bazy danych tylko podczas konfiguracji systemu. Po skonfigurowaniu maszyn administratorzy, tak jak użytkownicy, nie potrzebują wiedzieć, które maszyny przechowują kopie baz danych — po prostu dokonują zmian w bazie danych, a Ubik troszczy się już o zaktualizowanie wszystkich kopii bazy danych.

2.11 Dostęp do DFS z NFS

IBM dostarcza specjalizowanych narzędzi pozwalających na autoryzowany dostęp do systemu plików DFS:

- NFS/DFS Authenticating Gateway for AIX;
- NFS/DFS Secure Gateway for Solaris.

Bez nich użytkownicy klienta NFS mogą korzystać tylko z nieautoryzowanego dostępu do danych w przestrzeni plików DFS.