OpenGFS Lustre i możliwości ich działania w obrębie klastra OpenSSI

Projekt z przedmiotu RSO.A

Paweł Sławomir Kłósek
P.Klosek@elka.pw.edu.pl

16 czerwca 2005

Spis treści

1 OpenGFS

OpenGFS jest systemem plików z księgowaniem, który umożliwia współdzielenie jednego urządzenia dyskowego przez kilka komputerów, w taki sposób, że kilka komputerów może korzystać z jednego systemu plików nie nadpisując sobie wzajemnie danych.

1.1 Macierz dyskowa

Współdzielenie urządzeń we wcześniejszych wersjach OpenGFS może się odbywać np. się za pomocą technologii Fibre Channel lub iSCSI. Jednak najważniejszym założeniem tego rozwiązania, jest to, aby wszystkie komputery widziały współdzielone urządzenie jako urządzenie w katalogu /dev o takiej samej nazwie. Można to osiągnąć poprzez współdzielenie dowolnego urządzenia dostępnego jako urządzenie blokowe i odpowiednie zamapowanie urządzenia w systemie plików /dev na każdym z komputerów (odpowiednia wirtualizacja nazw w /dev).

Zewnętrzny zasób dyskowy dzielony należy podzielić na trzy partycje

- ok. 4MB na przechowywanie informacji o macierzy (np. informacje konfiguracyjne)
- ok. 128MB na przechowywanie zewnętrznego dziennika (journal)
- reszta na przechowywanie danych dziennika wewnętrznego

1.2 Struktury danych

Dinodes (distributed inodes), są jednostkami organizacji służącymi do przechowywania metadanych (czasem również danych jeśli plik jest mały). Każdy dinode zajmuje jeden blok. Dinody są tworzone tylko wtedy jeśli są potrzebne.

Plik

Jeśli plik jest mniejszy niż rozmiar pliku minus 232 bajty, plik można przechować w jednym dinode-zie w formacie: —meta-dane—dane z pliku—

Jeśli jednak plik jest większy to dane w dinodzie są przechowywane w następujący sposób: —meta-dane—adres 1-go bloku danych—adres 2-go bloku danych—...— .

Istnieje jeszcze trzeci format adresowania, niebezpośredni zamiast adresów bloków z danymi możemy przechowywać adresy do bloków z adreasami (do bloków z adresami do bloków z adresami ...) do bloków z danymi (można się zagłębić do 8 razy).

Dzięki tym trybom adresowania możemy przechowywać bardzo duże ilości danych [tab 1.]

Resorce groups Dane z plików umieszczane są z reguły blisko siebie w tzw. resorce block-ch, aby zmniejszyć liczbę ruchów głowicy nad talerzami dysków.

Katalogi Jest kilka metod przechowywania katalogów

- Bezpośrednio w dinodach bez tablicy hashującej
- W dinodach z tablicą hashującą (hash CRC32)
- W dinodach z adresami bloków tablicy hashującej.

Warto nadmienić dane na dysku są przechowywane w formacie big endian.

1.3 Zabezpieczenia przed zagrożeniami nadpisywania plików (locks)

W rozproszonym systemie plików na maszynach wieloprocesorowych należy wprowadzić trzy typy blokad:

- Między komputerami - jeśli kilka komputerów chce skorzystać z jednego bloku/pliku.
- Między kilkoma procesami na jednym komputerze (wykonywana za pomocą oprogramowania G-lock).
- Wewnątrz jądra systemu jeśli kilka wątków chce skorzystać z jednej struktury danych.

Jeśli chodzi o punkt pierwszy, to istnieją dwie metody zabezpieczania przed nadpisywaniem plików jedna zcentralizowana wykonywana za pomocą serwera blokad (lock server), ale wtedy oczywiście nie ma tolerancji na uszkodzenie właśnie tego serwera blokad, i rozproszona wykonywana za pomocą oprogramowania OpenDLM (Open Distributed Lock Manager), przy którym są z kolei problemy z wydajnością.

Przy centralnym blokowaniu plików wykorzystywany jest protokół memexp, który jest podobny do protokołu Device Memory Export Protocol (DMEP) jednakże do

transmisji informacji zamiast szyny SCSI wykorzystywana jest sieć lokalna przy wykorzystaniu protokołu IP.

1.4 Zabezpieczenia przed awariami

Aby zabezpieczyć system przed awariami spowodowanymi przez awarię jednego z komputerów, każdy komputer ma swój osobny dziennik zdarzeń (journal). Dzienniki mogą być lokalne (czyli na partycji 3) lub zewnętrzne (czyli na partycji 2). Dzienniki są po to rozdzielone, aby uszkodzony komputer nie uszkodził innych wpisów do dziennika i aby po awarii komputera dało się szybko przeprowadzić roolback).

Założeniem OpenGFS jest to, że awaria jednego z komputerów nie uszkodzi metadanych (więc struktura systemu plików nie ulegnie uszkodzeniu, jednak same dane mogą pozostać miejscami uszkodzone).

Aby zabezpieczyć cały system przed awariami stosuje się również izolowanie źle działających komputerów oraz ich zdalny restart, w tym celu można wykorzystywać watchdog-i, zdealne wyłączenie prądu przy zasilaczach programowalnych lub odcięcie komputera od swich-a Fiber Channel.

2 Lustre 1.0

Lustre jest klastrowym systemem plików, który w odróżnieniu od OpenGFS nie bazuje na jednym komponencie na którym można przechowywać dane. Według założeń Lustre ma być obiektowym systemem plików, w którym obiektami są dyski, komputery, pliki, katalogi itd.

2.1 Komponenty systemu

System składa się z dwóch typów serwerów serwera danych i serwera meta-danych.

Serwer metadanych (MDS - metadata Server)

Serwer metadanych (MDS - metadata Server) Serwer meta-danych ma zadanie przechowywać całą strukturę systemu plików za wyjątkiem samych danych. Według producenta testy wydajnościowe wykazały, że nawet w bardzo dużych rozwiązaniach jeden serwer meta-danych nie stanowi wąskiego gardła. Na serwerach metadanych wykorzystywany jest zmodyfikowany system plików ext3.

Serwer danych (OST - Object Stograge Targets)

Służy do przechowywania plików (lub fragmentów plików - zależnie od pliku). Na serwerze danych znajduje się zwykła partycja ext3 (chociaż zgodnie z dokumentacją może to być również JFS, ReiserFS lub XFS). Na partycji znajdują się pliki o nazwach w postacinumerycznej np. „2454213” - które to są numerami obiektów w systemie plików.

Serwer-y LDAP

Konfiguracja systemu plików może być przechowywana zarówno na każdym kliencie, który chce skorzystać z Lustre, jak również na pojedynczym serwerze LDAP.

2.2 Granulacja blokad na pliki

Informacja o blokadach na pliki (file locks) jest utrzymywana na serwerach metadanych. Blokady te mogą mieć różną granulację (istnienie pliku, katalog, atrybut,

sektor). Dzięki temu można zwiększyć wydajność - w zwykłym ext3 jeśli n użytkowników chce utworzyć plik w katalogu /tmp, każdy z nich musi wykonać blokadę na /tmp, założyć plik i wykonać operację odblokownia, co może być niewątpliwie czasochłonne. W Lustre, możliwe jest zrobienie tego całej tej operacji bez blokowania odpowiedniego katalogu.

Zabezpieczenia przed awariami Serwer meta-danych W tej wersji Lustre w systemie znajduje się maksymalnie jeden serwer aktywny i jeden system backupowy - który przejmuje rolę aktywnego - gdy tylko serwer aktywny przestanie odpowiadać.

Serwer backupowy musi mieć możliwość odcięcia zasilania lub odcięcia serwera aktywnego od sieci, ponieważ gdyby okazało się, że w sieci znajdują się dwa serwery aktywne, mogło by to doprowadzić do katastrofy.

2.3 Serwer danych

System nie jest podatny na awarię serwerów danych, i będzie po awarii bądź po odłączeniu serwera danych zachowywał się dalej stabilnie. Z tym, że obiekty (pliki) przechowywane na tym serwerze przestaną być dostępne. W trakcie usuwania awarii, nowe obiekty będą tworzone na innych serwerach danych.

Serwery danych mogą być również redundantne, tzn. dwa serwery mogą być podpięte do jednej macierzy dyskowej i po awarii jednego zawsze ten drugi może pracować zamiast tego pierwszego.

3 OpenSSI - Cluster File System (CFS)

W klastrze OpenSSI przyjęto założenie, że wszystkie zamountowane zasoby (łącznie z devfs i procfs) będą dostępne w formie plików - do których można się dostać przy wykorzystaniu dowolnego komputera w sieci lokalnej. Domyślnie w OpenSSI do spełnienia tych założeń wykorzystywany jest Cluster File System - rozwijany włącznie do projektu OpenSSI.

CFS jest wirtualnym systemem plików wykorzystywanym do współdzielenia zasobów dyskowych w klastrach OpenSSI.

CFS jest przezroczysty jeżeli chodzi o podmounowywanie zasobów, tzn. jeśli zamountujemy obsługiwany przez CFS fizyczny system plików (obecnie są to ext2,ext3,jfs) po zamountowaniu jest on przezroczyście widoczny na wszystkich komputerach w klastrze. Dodatkowym wariantem CFS-a jest High Availability - Cluster File System (HA-CFS). HA-CFS jest systemem plików opartym o macierz dyskową, i przy awarii jednego z komputerów podłączonych do macierzy macierz automatycznie przejmowana jest przez inny podłączony do niej komputer. Jest to rozwiązanie dużo prostsze od OpenGFS, gdzie wszystkie komputery podłączone do macierzy mogły z niej jednocześnie korzystać. To rozwiązanie doskonale może się sprawdzać w klastrach którym głównym przeznaczeniem nie jest szybkość dostępu do danych ale np. moc obliczeniowa - a dostęp do wspólnych danych jest jedynie dodatkiem.

Klaster OpenSSI może być jednocześnie klasterm przechowującym dane Lustre'a jednak jedynym rozsądnym rozwiązaniem przy takiej konfiguracji jest korzystanie z lokalnych nie udostępnianych przez sieć przez CFS systemów plików do przechowywania obiektów Lustre'a.

Jeśli chodzi o korzystanie z zasobów Lustre'a. Mountowanie jest propagowane na wszystkie komputery w klastrze. Można nadmienić również, że cały OpenSSI jest

dobrze zintegorowany z Lustre'm np. przy migracji procesów blokady na plikach przenoszone.

Na klastrze OpenSSI da się również korzystać z danych pochodzących z OpenGFS i Sistina GFS. W tej chwili da się przeprowadzić mountowanie zasobów OpenGFS jako dowolną partycję w systemie.

4 Wnioski

Opisane przeze mnie klastrowe systemy plików można podzielić na w zasadzie na 2 kategorie - z centralnym punktem przechowywania informacji OpenGFS i HA-CFS oraz CFS (w CFS mimo, że dane mogą być mountowane na wielu komputerach w klastrze to jednak dla każdego z tych zasobów to dane są przechowywane na jednym serwerze z ewentualnym serwerem awaryjnym). - system plików całkowicie rozproszony Lustre, gdzie dane są podzielone na obiekty i są przechowywane na wielu serwerach danych.

W zależności od zastosowań można wybrać jedną z konfiguracji: - CFS - dla klastrów nie przeznaczonych na przechowywanie danych - OpenGFS - dla klastrów w których szybkość dostępu do danych jest wartością krytyczną - Lustre - dla klastrów nastawionych na przechowywanie dużej ilości relatywnie szybko dostępnych danych.

Literatura

[2] *http://sourceforge.net/projects/opendlm*.

[3] *http://www.opengfs.org.*

[4] *http://www.lustre.org.*

[5] *http://www.sistina.com/products*

[6] openssi.org.

[7] http://sourceforge.net/projects/ssic-linux/.

[Wal05] Bruce J. Walker. Open Single System Image (openSSI) Linux Cluster Project. *2005.*