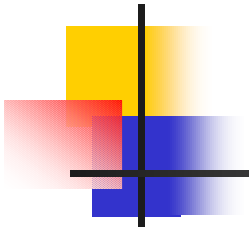


클러스터 파일 시스템

Networking Laboratory

요약: 배희정

2002. 1. 22

SKKU

1. 서론

Networking Laboratory

1. 클러스터 시스템
   1. 컴퓨터들을 네트워크로 연결하여 사용자에게 단일이미지를 제공하는 시스템
   2. 고확장성, 고가용성, 가격대 성능비 뛰어남
   3. DAS(Direct Attached Storage)구조

SKKU

1. 서론

Networking Laboratory

2. 네트워크 저장장치
   1. 고속의 내트워크로 저장장치 연결. 여러노드들이 공유
   2. High Scalability, High Availability, High Speed
   3. 자료 공유 ✍ 특정 서버의 병목현상 제거
   4. SAN
      1. 파이버채널로 대변되는 저장장치 전용네트워크
      2. 네트워크로 연결되는 저장장치 구조

SKKU

1. 서론

Networking Laboratory

3. SAN 가상화
   1. 사용자 또는 상위계층 응용에 하드웨어 투명성을 지원할 수 있는 일관된 기능의 인터페이스를 제공하는 것.
   2. Virtual Memory 개념과 비슷
   3. 소프트웨어로 제공
   4. 기능: 소프트웨어 RAID, Snapshot, Storage on Demand

SKKU

1. 서론

Networking Laboratory

4. 클러스터 파일 시스템.
   1. SAN virtualization을 기반으로 함.
   2. 대용량 및 여러 호스트에서 파일을 공유할 수 있는 특성.
   3. 고려사항
      1. 64비트용 파일시스템
      2. 파일공유 시멘틱의 지원 및 이에 따른 효율적인 메타데이터 관리
      3. 빠른 회복기능
      4. 대량의 파일 저장을 위한 디렉터리 관리기법

SKKU

2. 관련연구

Networking Laboratory

1. xFS, Frangipani 파일시스템
   1. 분산공유 파일시스템
   2. DAS 구조에서 동작 ✍ 변형요구됨
2. Pool Driver: 볼륨관리
   1. SAN환경에서 수행
   2. 레이드 레벨로서 striping과 mirroring만을 지원
   3. Snapshot이나 데이터의 동적이동은 지원못함
   4. 파일공유 시 성능저하(파일캐쉬 지원 안 하므로)
   5. 유연성 떨어짐 (저장장치 기반 잠금 사용)

SKKU

2. 관련연구

Networking Laboratory

3. SANergy
   1. NFS, CIFS와 결합하여 SAN에서 데이터 공유 지원
   2. 이기종 운영체제에서도 수행
   3. 중앙 집중된 논리 볼륨 관리자에 의존하는 단점

SKKU

3. 시스템환경

Networking Laboratory

- SANtopia
  - 각 호스트에서 수행되는 Linux운영체제에 포함된 시스템 소프트웨어
  - 논리볼륨 지원
  - 대용량 파일시스템 지원
  - 호스트사이의 통신수단으로는 LAN사용

SKKU

4. 볼륨 관리

Networking Laboratory

- 기능: 물리적으로 독립적인 여러 개의 저장공간을 모아 하나의 가상적 기능을 제공하는 논리 저장공간을 상위 계층에 제공하는 것.
- 볼륨 관리자는 운영체제 커널의 가상 디바이스 드라이버로 동작.
- 하나의 논리볼륨은 하나의 가상디바이스 파일로 표현.

SKKU

4.1 볼륨구성

Networking Laboratory

- **논리볼륨** – 디스크 파티션을 최소구성단위로 하는 연속적인 주소공간을 이루는 확장가능한 디스크 파티션들의 집합.
- **논리볼륨 구성시에 고려되어야 하는 요소**
  - 익스텐트 크기
    - 동일크기를 갖는 연속적인 공간.
    - 정보의 저장을 위해 할당될 수 있는 디스크공간의 최소단위
    - 익스텐트 크기는 하나의 논리 볼륨에 대해 동일.
  - 메타데이타
    - 논리볼륨 구성에 대한 정보 및 해당 논리볼륨을 관리.
    - 디스크 파티션의 헤더 영역에 기록.

SKKU

4.2 Software RAID

1. CPU가 RAID관리
2. 추가로 일렬로 디스크장치를 연결한 형태의 concatenation 지원

- Hardware RAID: RAID 컨트롤러가 제어.
  - CPU 활용도 증가, 버스 트래픽 감소
  - 고가.
  - 컨트롤러의 오류발생시 모든 RAID 시스템 사용 불가

Networking Laboratory
SKKU

4.3 Mapping

Networking Laboratory

1. 물리적 공간에 대한 가상화
   - 논리저장공간과 물리저장공간의 매핑과정에 이루어진다.
2. 논리주소를 해당 물리주소로의 매핑에 있어서 가변성 수용 – 매핑 테이블 유지
3. Snapshot
   1. 백업 효율적으로 처리
   2. 매핑관리자는 스냅샷 요구시 매핑테이블 복제,유지.
   3. 데이터 변경시 – 변경시 복제기법(copy-on-write)
      기존 매핑테이블 유지하면서 변경된 데이터는 새로운 매핑 테이블이 관리. 백업 시 기존 테이블 사용.

SKKU

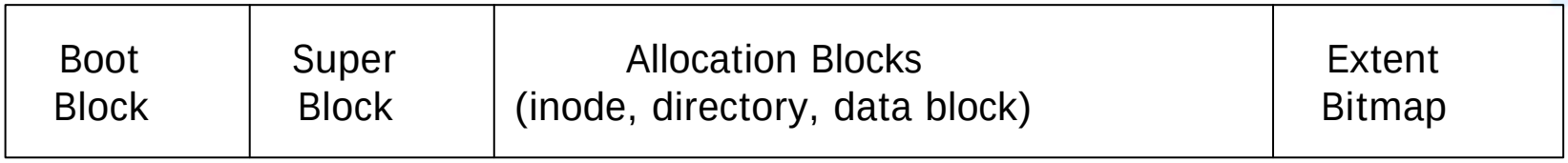
5. 파일시스템

Networking Laboratory

- SANtopia의 파일 시스템
  - SAN환경기반, 클러스터 파일시스템
  - 64bit address 지원
  - 익스텐트(1KB~64KB)를 할당의 기본단위로 사용
  - 대등관계모델의 공유파일시스템

SKKU

5.1. 파일시스템 레이아웃



- Super Block: 파일시스템에 대한 정보 저장
- 비트맵 Block: 각 익스텐트의 사용여부를 표시하여 익스텐트의 할당, 회수관리
- 할당 블록(Allocation Blocks): 데이터, inode, 디렉터리를 저장.

5.2. 파일시스템의 생성

- SANtopia 파일시스템 : 논리볼륨관리자에 의해 제공되는 논리 볼륨 위에 생성
- 여러 개의 물리디스크를 바탕으로 스트라이핑 또는 연결형 형식 등으로 구성되는 논리볼륨과 그 위에 파일시스템이 존재하게 된다.

Networking Laboratory
SKKU

5.3. 메타데이타 관리

1. 동적 다단계 inode
   - 대용량 파일시스템을 위한 새로운 메타데이타 구조로 대용량 파일을 저장하기 위함
2. 다단계 동적 확장 해쉬
   - 다수의 파일이 존재 가능한 디렉터리 관리 구조

Networking Laboratory
SKKU

5.3.1. 다단계 inode구조

- Stuffed inode 기법 – 익스텐트의 사용으로 인한 메모리의 낭비를 줄이기 위한 방법. inode를 초과하기 전까지 inode 내에 데이터를 저장함
- Dynamic multi-level inode

Networking Laboratory
SKKU

5.3.2. 디렉터리 관리

- 다수의 파일 처리
  - B+트리 사용하는 방법 – xFS
  - 확장 해쉬 기법 - GFS

Networking Laboratory
SKKU

5.4. 저널링(Journaling)

- 파일시스템에 대한 연산 발생시 메타데이타에 대한 변경정보를 로그에 기록하여, 오류 발생시 로그에 기록된 메타데이타를 이용하여 복구하는 방법.

Networking Laboratory
SKKU

5.5. 파일캐쉬

Networking Laboratory

1. 상호 협력 캐쉬 기법
    1. 요청된 파일이 자신의 버퍼에 존재하지 않는 경우, 디스크로부터 읽지 않고 다른 호스트의 버퍼에 존재하는가 먼저 검색.

SKKU

5.5. 파일캐쉬

Networking Laboratory

2. 전역버퍼리스트 관리
   1. 각 호스트의 파일 캐쉬에 대한 정보 취합
   2. 특정호스트에 의해 관리, 여러 호스트가 나누어 관리(부하 분할)
   3. 버퍼관리자와 잠금 관리자 통합
      - 로컬버퍼 변경 때마다 전역리스트 갱신 시 성능저하
      - 모든 파일에 대한 연산을 처리하기 위해서는 먼저 파일에 대한 잠금을 얻어야 한다.
   4. 콜백 잠금 기법
      - 전역잠금 관리 기법
      - 잠금의 할당이 특정서버에 의해 관리
      - 특정 호스트가 잠금 획득 시 사용 후 반납 안 해도 됨
      - 문제 & 해결방안
        - Starvation – 예약 시멘틱 제공
        - 오류 취약성 – 잠금서버 중복

SKKU

5.5. 파일캐쉬

Networking Laboratory

1. 전역 호스트
   - 전체 파일시스템에 대한 정보 관리
   - 잠금 서버와 버퍼 서버의 역할 동시담당
   - GLM모듈 – 전역 잠금 리스트
     LLM 모듈 – 개별 잠금 리스트
     - GLM과 LLM은 항상 일치 (일관성 깨지는 문제발생)
     - GBM과 LBM은 내용차이 있음(오버헤드 줄이기 위해)

SKKU

6. 시스템 관리

Networking Laboratory

- 중앙 집중의 강력한 관리기능 요구
  - 성능 모니터링
  - 오류 탐지 기능
  - GUI기반 구성 관리 기능

SKKU

7. 관련응용

Networking Laboratory

- 고차원 색인 구조 – 병렬 환경 이용
- 공간 데이터 베이스 관리 시스템

SKKU