Managing Over 1,000 OpenSearch Clusters in a Private Cloud

Sun Ro Lee, LINE

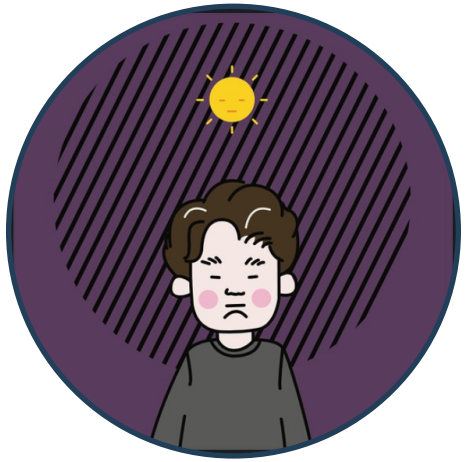
LINE

Contents

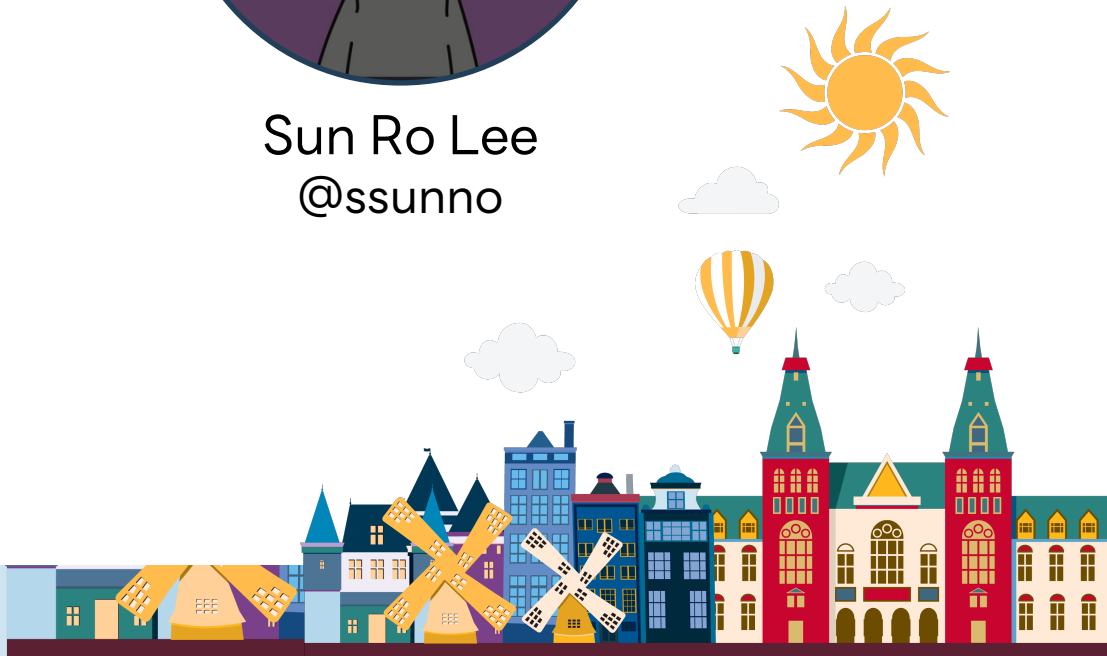
OpenSearch

Introduction

- Cloud Service Engineer at LINE
- OpenSearch service 개발과 운영 담당
- LINE에서 5년 동안 OpenSearch 관련 업무



Sun Ro Lee
@ssunno



Private cloud 를 운영하는 이유

- 보안과 데이터 보호
  - LINE은 메신저 기반 플랫폼으로 다양한 사용자 데이터를 처리하고 있으며 개인정보를 비롯한 사용자 데이터에 높은 수준의 데이터 보안 규정 준수가 필요함
- 비용 효율성
  - 대규모 서비스를 운영할 때 프라이빗 클라우드는 퍼블릭 클라우드를 이용하는 것 보다 비용이 저렴할 수 있음
- 기능 확장성 및 유연성
  - 서비스의 특별한 요구사항에 맞춘 기능이나 하드웨어 구성을 제공할 수 있음.

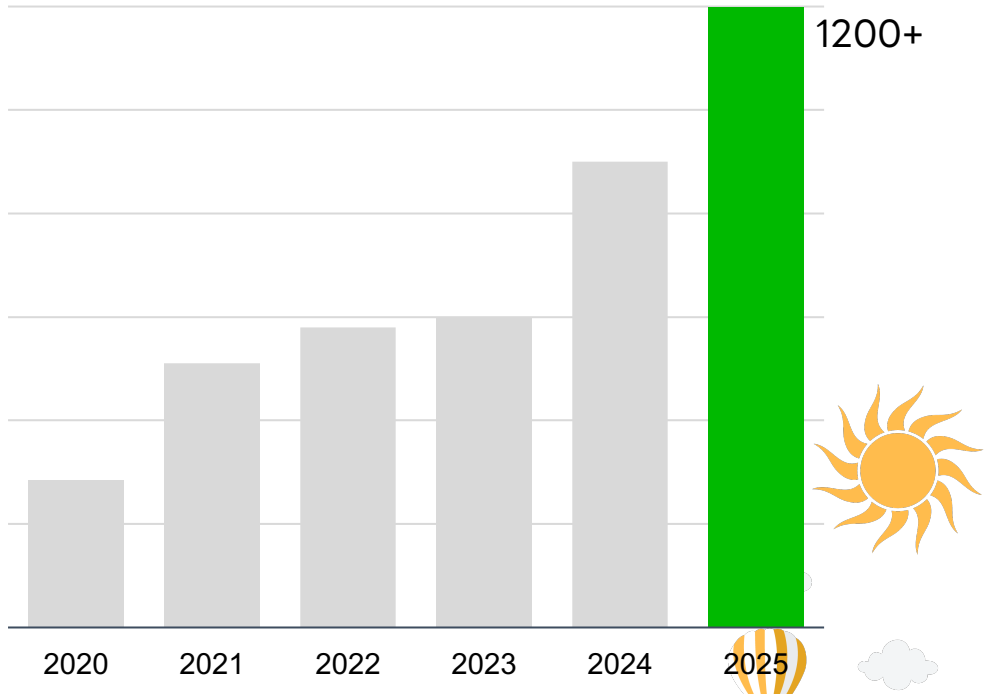
OpenSearch on LINE

OpenSearch

운영 규모

- 클러스터 수: 1,200+
- 노드 수: 10,000+
- 데이터 규모: 6PB+
- 리전: 2개 / zone: 6개 에서 서비스 중
- 사용 목적

  로그 저장소, 데이터 분석, 실시간 모니터링
  검색 엔진, 데이터 파이프라인..



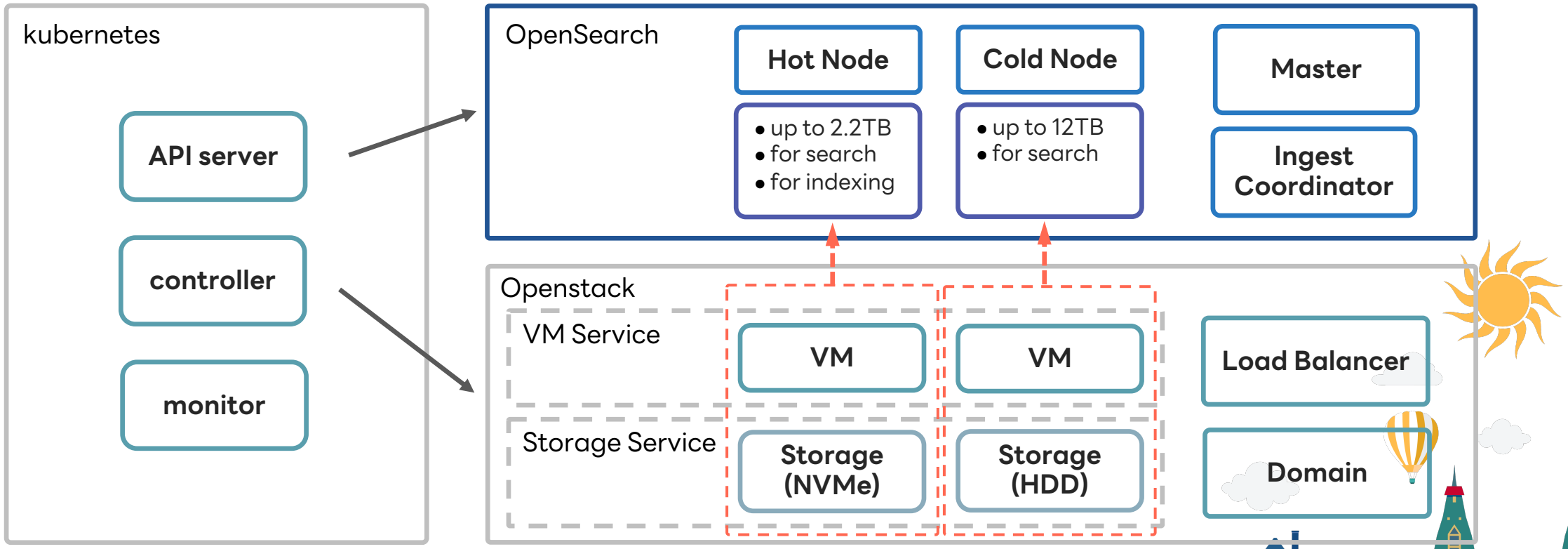
OpenSearch

OpenSearch with Network Storage

LINE

OpenSearch with Network Storage

OpenSearch Architecture



OpenSearch with Network Storage

왜 OpenStack VM 을 사용할까요?

- 쿠버네티스는 서비스를 일관성 있게 유지할 수 있는 장점이 있지만, 컨트롤 플레인 고장 (API server, etcd 등) 으로 모든 OpenSearch 클러스터가 영향을 받을 수 있는 문제가 있음.
  - 예시: chatGPT 에서 k8s API 서버 과부하로 인한 서비스 대규모 장애가 있었음
  - 컨트롤플레인이 고장인 상황에서도 OpenSearch 서비스를 사용할 수 있도록 보장하기 위해 VM 으로 구조를 변경함
- OpenSearch 클러스터들을 서로 다른 VPC에 구축하려는 수요가 있음.

OpenSearch with Network Storage

Network Storage 도입 배경

- OpenSearch에서는 local storage를 사용할 것을 권장하고 있음.

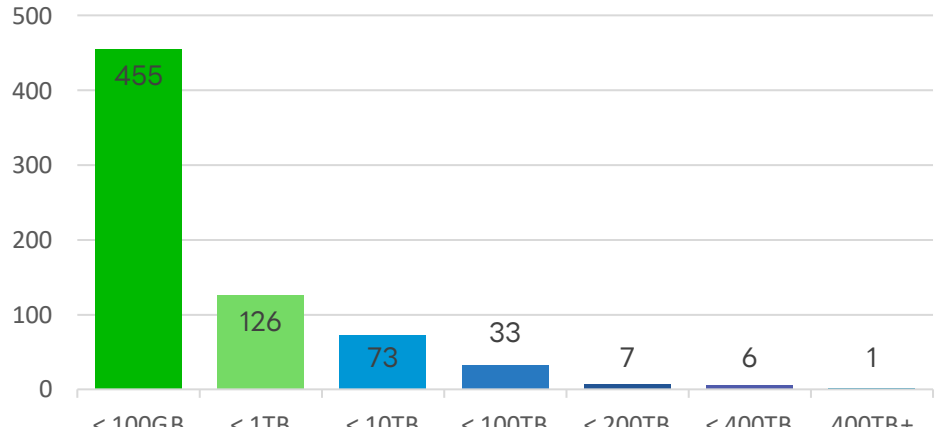
> **File system recommendations**
>
> Avoid using a network file system for node storage in a production workflow. Using a network file system for node storage can cause performance issues in your cluster due to factors such as network conditions (like latency or limited throughput) or read/write speeds. You should use solid-state drives (SSDs) installed on the host for node storage where possible.

- 최초에는 대형 SSD를 탑재한 서버로 OpenSearch를 운영했음
- 그러나 OpenSearch 전용 서버를 운용하는 것은 번거로운 일이고 대형 클러스터를 구축할 때에는 node affinity 문제가 생길 수 있음.

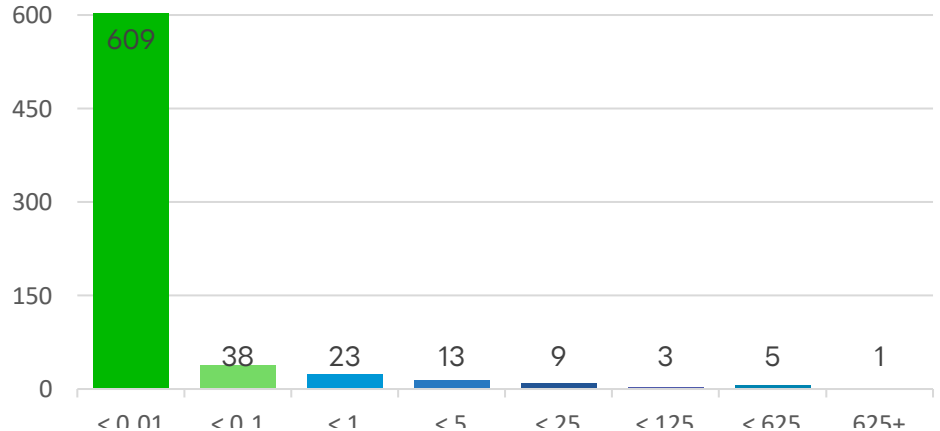
OpenSearch with Network Storage

Network Storage 도입 배경

- 고성능 노드와 대규모 데이터 저장 수요 등 요구사항이 다양해지면서 여러 타입의 서버가 필요하게 되었음.
- 서버 관리의 용이함과 다양한 노드 티어를 지원하기 위해서 network storage를 도입
- OpenSearch 에서 local storage를 권장하는 이유는 성능 문제 때문
  ▶ OpenSearch에서는 높은 디스크 성능이 중요하기 때문에, NVMe over Fabric 사용



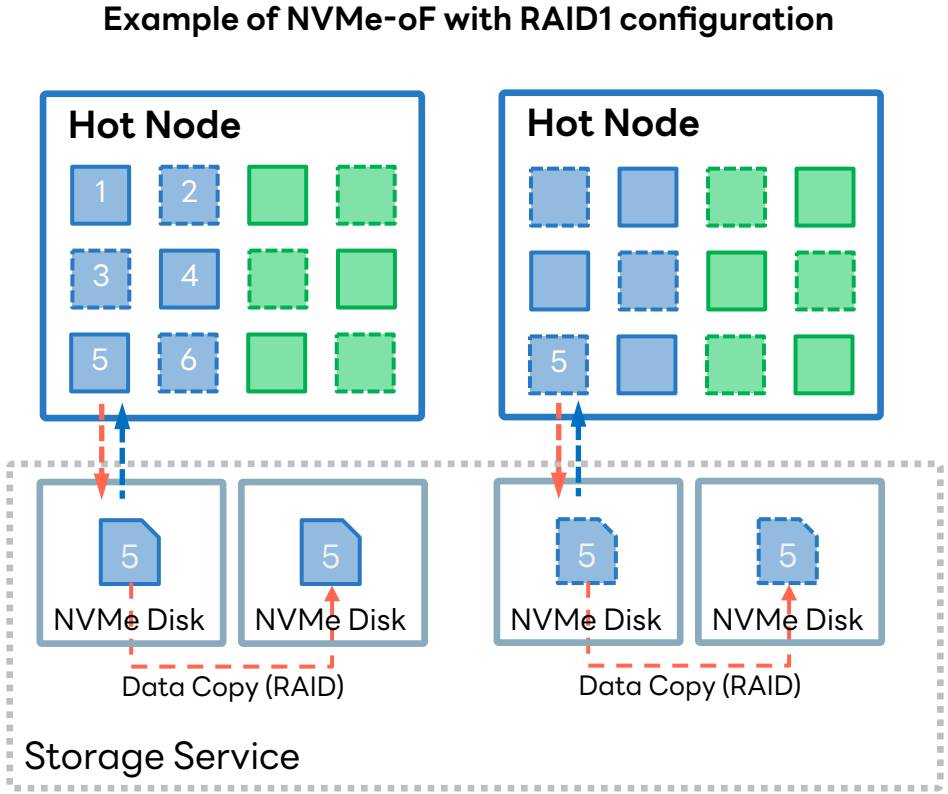
Average Stored data distribution of clusters



Average QPS(query per second) distribution of clusters

NVMe over Fabric

- NVMe over Fabric(NVMe-oF) 는 NVMe 디스크를 네트워크를 통해 연결할 수 있게 해주는 네트워크 프로토콜
- RAID 를 구성할 수 있어 데이터 보호와 내결함성을 갖춘 디스크를 제공할 수 있음
- 장점: 낮은 지연시간과 병렬처리 기능으로 높은 성능과 확장성을 제공할 수 있음
- 단점: NVMe 스토리지 서버는 다른 타입보다 서버 유지비용이 더 높다.

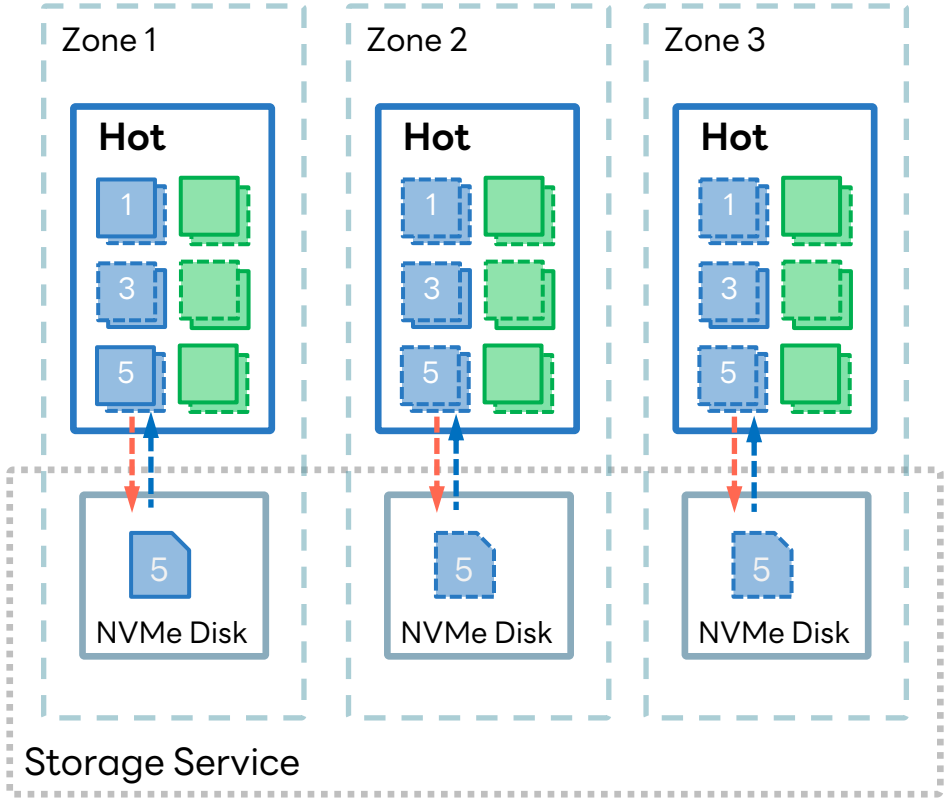


OpenSearch with Network Storage

NVMe over Fabric - Lite Storage

- NVMe-oF 에서 **RAID를 사용하지 않는** 디스크 타입.
- 장점: RAID1 구성에 비해 네트워크 대역폭과 스토리지를 절반만 사용하기 때문에 더 저렴한 비용으로 네트워크 스토리지를 이용할 수 있다.
- 단점: 디스크 고장 시 데이터를 잃어버릴 수 있다
  - Local Storage와 동일한 수준으로 데이터 관리
  - OpenSearch에서 multi-zone shard affinity, shard replication 설정으로 내결함성 확보
  - 여유 디스크를 더 많은 shard replication 에 사용할 수 있으므로 검색 성능에 리소스를 더 투자할 수 있음

Example of NVMe-oF without RAID configuration



NVMe over Fabric – 대규모 데이터 마이그레이션 사례

- OpenSearch 대규모 마이그레이션 케이스: 데이터플랫폼 조직에서 자체 운영하는 OpenSearch를 클라우드로 이전하는 작업 (이전 대상 데이터: 6.1PB)
- Hot-Warm architecture를 적용함 (Hot: 2.0PB, Cold: 4.1PB)
- NVMe-oF Lite Storage 적용 후 프로젝트 비용(서버, 스토리지 구매 비용) 22% 감소
  - 프로젝트 비용 중 NVMe-oF 서버는 33%
  - NVMe-oF 서버 비용은 기존 대비 53.8%로 감소 (Lite Storage 적용)

OpenSearch

Case Study: Instability in OpenSearch with Network Storage

LINE

Shard failed with “Checksum Failed” error

- Network 디스크에 저장된 샤드 중 일부에서 ”checksum failed (hardware problem?)” 에러가 발생하는 문제
  - 문제가 발생한 샤드는 유효하지 않은 것으로 간주되어서 unassigned 상태가 됨.
- 문제는 Cold Storage(block storage volume) 에서 발생
  - NVMe 를 사용하는 Hot storage에서는 발생하지 않음
- OpenSearch가 replica 를 이용해서 샤드를 복구할 수 있었다.
  - ▶ 데이터를 쓸 때가 아니라 읽을 때 문제를 발견할 수 있으므로 장애 발생 시점을 알기가 어렵다.
  - ▶ Primary와 Replica 샤드가 서로 다른 데이터를 가지고 있는 상태가 장시간 지속 된다. (한 쪽이 체크섬 불일치 상황일 때)

OpenSearch with Network Storage

OpenSearch

- 다양한 실험에도 불구하고 원인을 찾지 못했음.
  (장애 재현이 어려워서 시간이 많이 소요됨)
- 다양한 버전 조합으로 테스트를 진행했지만 직접적인 원인은 아님
- Block Storage 노드에서만 문제가 발생하는 것으로 확인됨.

: PASSED : FAILED

(network) Storage Type	HDD	SSD	NVMe	
OpenSearch Version	ES 7.10	2.4	2.11	2.15
Lucene Version	8.7.0	9.4.2	9.7.0	
OS Version	CentOS 7.9	Rocky Linux 8.9		
Kernel Version	3.10.0-1160.49.1	4.18.0-513.5.1	6.1.54-1.20230921	
Memory Swap	ON	OFF		
Filesystem	EXT4	XFS	XFS (wsync,nobarrier)	

OpenSearch with Network Storage

- 추정 원인
  - OpenSearch 노드가 높은 부하 상황일 때 파일시스템 캐시가 디스크에 제대로 기록 되지 않음
  - Block Storage(Ceph) 의 부하에 따라 달라지는 지연 시간이 디스크 쓰기에 영향을 주고 있음
  - 인덱싱 도중에 Node VM 에서 수십GB의 파일시스템 캐시가 유지되고 있음
  - Lucene은 트랜잭션을 지원하지 않으므로 데이터 쓰기 과정에서 오류가 발생하더라도 알 수 없음
    → 데이터를 읽을 때 체크섬 에러가 발견됨
- Pain point
  - 높은 부하 상황에서만 가끔 문제가 발생하고 데이터를 읽을 때 문제를 발견할 수 있기 때문에 문제를 재현하기 어렵다.

OpenSearch with Network Storage

OpenSearch

- 결정적인 원인
  - OpenSearch가 서비스 되고 있는 2개의 Openstack 리전 중 한 곳에서만 발생함
  - Openstack Nova에서 사용하는 디스크 캐시 옵션(disk_cachemods)이 writeback 으로 설정된 리전에서만 문제가 발생하고 있었음
  - writeback 옵션은 캐시를 사용해서 디스크 성능을 높이지만 손실 위험이 있음

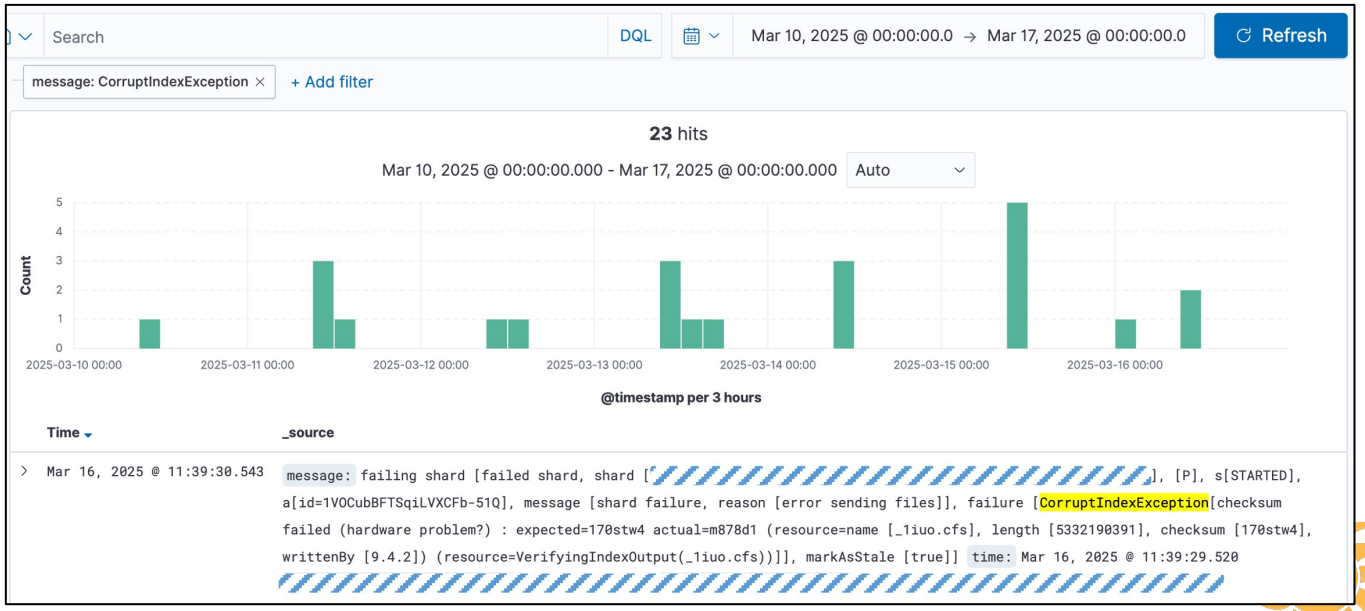
Because the host page cache is enabled in this mode, the read performance for applications running in the guest is generally better. However, the write performance might be reduced because the disk write cache is disabled.

- writeback: With caching set to writeback mode, both the host page cache and the disk write cache are enabled for the guest. Because of this, the I/O performance for applications running in the guest is good, but the data is not protected in a power failure. As a result, this caching mode is recommended only for temporary data where potential data loss is not a concern. NOTE: Certain backend disk mechanisms may provide safe writeback cache semantics. Specifically those that bypass the host page cache, such as QEMU's integrated RBD driver. Ceph documentation recommends setting this to writeback for maximum performance while maintaining data safety.
- directsync: Like "writethrough", but it bypasses the host page cache.
- unsafe: Caching mode of unsafe ignores cache transfer operations completely. As its name implies, this caching mode

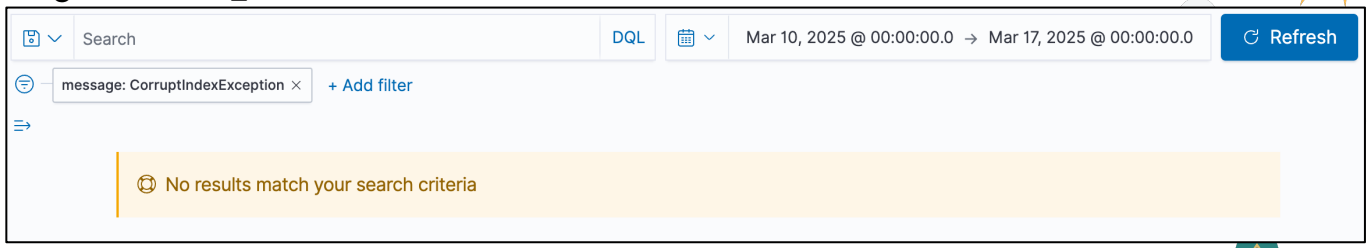
OpenSearch with Network Storage

- `CorruptIndexException[checksum failed (hardware problem?)` 에러가 발생하고 샤드는 unassigned 상태로 변함 (데이터 탈락)
- writeback 으로 설정된 리전에서만 문제가 발생하고 있음
- 데이터 사용 통계 (cold storage)
  - Region A : 600 TB +
  - Region B : 2000TB +

Region A. disk_cachemods=writeback



Region B. disk_cachemods=none



LINE

writeback VS none

- disk_cachemods=none 으로 설정된 리전에서는 문제가 발생하지 않았음
- writeback 은 더 빠른 I/O 성능을 제공할 수 있다고 설명하고 있지만 실 사용 성능은 비슷하다

		Sequence Read, Write				4K Random IOPS		
		none	writeback	none/wb		none	writeback	none/wb
1 Jobs	Read (MB/s)	234	406	58%	Read (OPS)	4.9K	5.0K	100%
	Write (MB/s)	100	385	27%	Write (OPS)	5.0K	5.0K	100%
4 Jobs	Read (MB/s)	395	493	81%	Read (OPS)	5.0K	5.0K	100%
	Write (MB/s)	351	510	69%	Write (OPS)	5.0K	5.0K	100%
32 Jobs	Read (MB/s)	**509**	**512**	**100%**	Read (OPS)	**5.0K**	**5.0K**	**100%**
	Write (MB/s)	**511**	**512**	**100%**	Write (OPS)	**5.0K**	**5.0K**	**100%**

* Random IOPS 는 5K 로 QoS 가 설정되어 있음.

- Block Storage 를 사용하려면 디스크 캐시 설정에 유의

Ceph Storage health check failed

- 하나의 리전에서 같은 ceph를 사용하는 노드 대부분이 다운되는 대규모 장애
- 노드에서 'disk health check failed' 에러와 함께 노드가 클러스터에서 탈락한다.
- Ceph 로 구성된 스토리지에서만 문제가 발생함
- 문제가 심각한 이유
  - OpenSearch 안에서 발생한 문제가 아니며 여러 클러스터가 동시에 영향을 받는다.
  - 여러 개의 노드가 다운되기 때문에 Primary, Replica 샤드가 동시에 유실될 수 있다.
  - 영향을 받는 모든 OpenSearch가 데이터 복구를 시도하기 때문에 부하가 급증한다.
  - ▶ 리전 전체에 문제가 생기고 클러스터가 스스로 문제를 해결하지 못하는 상태

OpenSearch with Network Storage

원인 분석

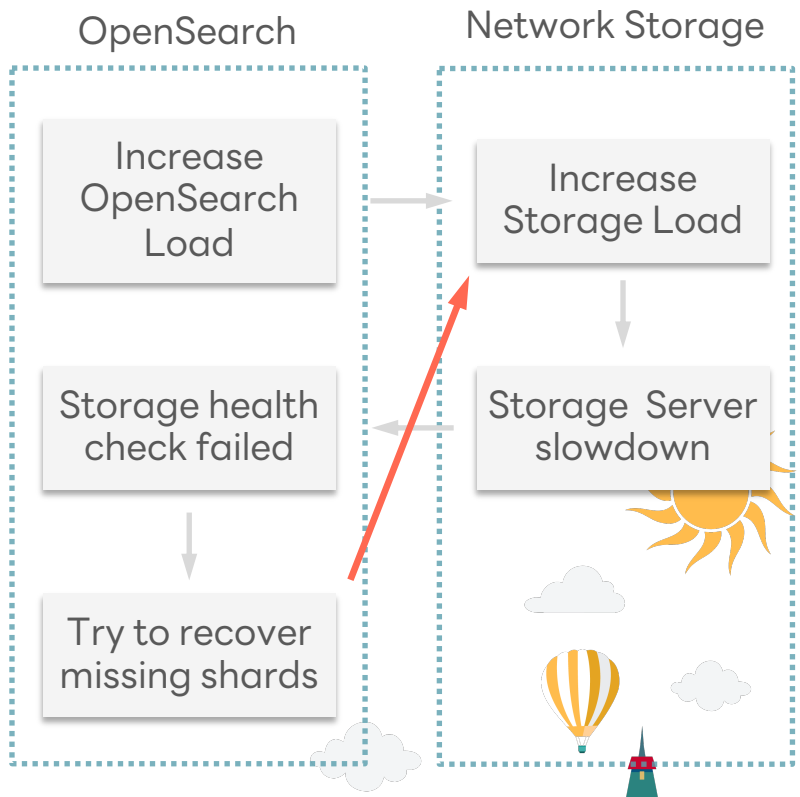
01 block storage는 데이터를 블록 단위로 나눠서 저장하는 구조로, 여러 개의 스토리지 서버가 데이터를 나눠서 처리한다.

- 01-A 따라서 네트워크 스토리지에 가해지는 부하가 증가하면 ceph 클러스터(스토리지 서버) 전반적으로 부하가 늘어날 수 있다.
- 01-B 스토리지 서버 내에 임의의 볼륨이 자신이 부하를 일으키고 있지 않아도 성능과 응답성이 떨어질 수 있다.

02 OpenSearch 노드는 볼륨 헬스체크에 실패하면 클러스터에서 탈락한다.

03 네트워크 스토리지를 사용하는 여러 개의 OpenSearch 노드가 헬스체크 문제로 동시에 탈락할 수 있다.

04 OpenSearch는 스스로 데이터를 복구하려고 시도 하고, 이 때 부하가 더 증가한다.



OpenSearch with Network Storage

OpenSearch

How to fix ? - Recover from disaster

01 네트워크 스토리지를 사용하는 "**모든**" OpenSearch 클러스터의 노드간 이동 속도를 낮춘다.
- ▶ indices.recovery.max_bytes_per_sec

02 '동시에 이동하는 샤드 수' 제한을 높여서 부하를 줄인다.
- ▶ cluster.routing.allocation.node_concurrent_recoveries
- ▶ cluster.routing.allocation.node_initial_primaries_recoveries

03 네트워크 스토리지가 안정적인 상태가 될 때 까지 불필요한 샤드 이동을 중단시킨다
- ▶ cluster.routing.rebalance.enable

04 Ceph 클러스터와 네트워크 노드가 안정되면 부하를 조금씩 증가시킨다

OpenSearch with Network Storage

How to fix ? - Preventing future outage

- OpenSearch 전용 서버 운영에 어려움이 없다면 로컬 스토리지를 사용하는 것이 더 좋다.
  - Tiered node를 구성하려면 여러 종류의 서버가 필요하다.
  - 서버 수가 적다면 affinity 가 문제가 될 수도 있다.
- 충분히 큰 네트워크 스토리지 서비스가 필요하다. (많은 비용이 필요할 수 있음)
  - 퍼블릭 클라우드를 사용하는 것이 (비용 측면에서) 더 저렴할 수도 있다.
  - Searchable Snapshot이 더 나을 수도 있다. (disk health check 를 하지 않기 때문에 노드 장애로 이어지지 않는다.)

OpenSearch with Network Storage

OpenSearch

Benefit of Network Storage

- 디스크 고장 내결함성과 서버 고장 복구시간 감소
  - 스토리지 수준에서 디스크 고장 복구 기능이 있으므로 디스크 고장을 고민하지 않아도 된다.
  - 컴퓨팅 노드 (VM)가 고장났을 때, 데이터를 복구하는 대신 VM을 새로 만드는 것으로 더 빠르게 복구가 가능하다.
    - 예시: 8TB 데이터를 가진 cold node의 데이터를 복구하는 데 노드 복구 속도를 200MB/s 로 설정하면 41,943초(=11.65시간) 가 필요하다.
- OpenSearch 전용 서버 관리 부담이 줄어듦
- 대형 클러스터 구축 시 shard affinity 설정이 용이함

OpenSearch with Network Storage

Next Plan

- 10PB + 규모로 데이터 증가가 예정되어 있음
  - 클러스터와 노드 수가 늘어남에 따라 노드 유지·보수를 자동화 할 계획
- Petabyte scale OpenSearch
  - 더 큰 규모의 단일 클러스터를 지원하기 위한 기능 추가
    - 대형 클러스터에서 발생하는 마스터 노드 지연 등, 알려진 문제들을 해결하는 과제가 있음
  - Searchable Snapshot 에서 알려진 버그를 수정하고 Searchable Snapshot 데이터 lifecycle 자동 관리 기능을 추가할 예정

OpenSearch

Thank you :)

LINE