Santeri Tallqvist

Ceph VMwaren tallennusalustana

Metropolia Ammattikorkeakoulu

Insinööri (AMK)

Tieto- ja viestintätekniikan tutkinto-ohjelma

Insinöörityö

17.05.2019

Metropolia

Tekijä Otsikko Sivumäärä Aika	Santeri Tallqvist Ceph VMwaren tallennusalustana 27 sivua 17.05.2019
Tutkinto	insinööri (AMK)
Tutkinto-ohjelma	tieto- ja viestintätekniikka
Ammatillinen pääaine	Communication Networks and Applications
Ohjaajat	Tapio Wikström

Tämä opinnäytetyö käsittelee suorituskykymittauksien tuloksia siitä, miten Ceph, ohjelmistopohjainen tallenusympäristö toimii Vmwaren tallennusalustana käytössä olevan Hitachin block storage -ympäristön sijaan. Tämä tutkimus tehtiin CSC – Tieteen Tietotekniikan Keskus Oy:lle kesällä 2017 minun työharjoitteluni osana.

Ceph on vapaan lähdekoodin tallennusjärjestelmä, joka hyödyntää objektipohjaista tallennusta. Tätä objektidataa voidaan hyödyntää Cephin työkalujen kuten RADOS Block Devicen ja CephFS:n lävitse esimerkiksi lohkotallennustilana tai tavallisena tiedostojärjestelmänä. Ceph mahdollistaa minkä vain laitteiston käytön ja on täten halvempi kuin suurilta palvelintarjoajilta valmiina ostetut palvelimet. Tämä tekee Cephistä houkuttelevan vaihtoehdon yrityksille.

Vmware on yritys, joka tarjoaa monia erilaisia palveluita, joista opinnäytetyöhön keskeisiä on sen virtualisointiin liittyvät palvelut, kuten itse virtuaalikoneista vastaava ESX ja vSphere. Tarkoitus on tutkia, miten Cephiä voidaan käyttää näiden osien tallennuspohjana. Cephiä käytetään jo nyt CSC:llä toisen virtualisointiohjelmiston OpenStackin tallennusalustana, ja se on toiminnaltaan hyvin samankaltainen kuin juuri edellä mainitut ESX ja vSphere, joten halusimme nähdä, miten Ceph toimii Vmwaren kanssa.

Suorituskykymittauksissa Ceph oli huomattavasti jo käytössä olevaa Hitachin levyjärjestelmää heikompi. Ceph on kuitenkin päivittynyt, kehittänyt suorituskykyään ja on edelleen suuressa osassa CSC:n kehitystä tulevaisuudessa. Ceph tulee olemaan tallennusalustana CSC:n uudelle supertietokoneelle.

Avainsanat	Ceph, VMware

Author Title Number of Pages Date	Santeri Tallqvist Ceph as VMware's storage back-end 27 pages 17 May 2019
Degree	Bachelor of Engineering
Degree Programme	Information and Communication Technology
Professional Major	Communication Networks and Applications
Instructors	Tapio Wikström

This thesis presents and discusses the results of testing how Ceph, a software based storage system works as a back-end for Vmware virtual machines instead of the currently used Hitachi Block Storage. This thesis was done as research for CSC – IT Center for Science Ltd where the author of the thesis did his internship during summer of 2017.

Ceph is a new free and open source storage system that utilizes object storage. Through gateways like RADOS Block Device and CephFS it can be used as block storage to replace other block storage systems or just as a regular file system. It enables customers to use any commodity hardware to set up a storage environment and is for this reason a very tempting solution for companies to use.

Vmware is a company offering many solutions especially on virtualization with many of its different parts. In this thesis the focus was on its hypervisor ESX and its controlling platform vSphere to see how they run with Ceph provided storage. Ceph is already used at CSC predominantly as a back-end for OpenStack, which is similar in how it works compared to ESX and vSphere; thus there was a need to see how it would run as back-end for other services.

The test results showed that Ceph was still lacking much in performance compared to the Hitachi Block Storage system that was already used by Vmware.

Ceph, however, has received updates to improve its performance and usage at CSC is ever increasing. It will be used as a storage back-end for Finland's new super computer.

Keywords	Ceph, Vmware

Sisällys

Lyhenteet

ESX	Elastic Sky X. Vmwaren itse virtualisoinnista vastaava palvelinkone.
KVM	Kernel Virtual Machine, virtuaalikone suoraan Linuxissa.
OSD	Object Storage Device on osa laitetta, johon data tallennetaan.
POSIX	Portable Operating System Interface. Standardi, jota Unix-tyyppiset käyttöjärjestelmät, kuten Linux seuraavat.
MDS	Metadata Server. Ceph-ympäristön osa, joka pitää yllä POSIX-ympäristön komentoja.
SAN	Storage Area Network, joka sisältää block-tallennustilaa

1 Johdanto

Tämä opinnäytetyö käy läpi sitä, miten hyvin Ceph toimii tallennusalustana Vmwaren virtuaaliohjelmistojen ja eritoten virtuaalikoneiden pohjaksi. Opinnäytetyö on tehty CSC – Tieteen Tietotekniikan Keskus Oy:lle tekemäni tutkimustyön pohjalta, joka ajoittui kesäharjoittelun ajalle kesälle 2017. Cephin ylläpito ja tämä tutkimus oli pääasiallinen työnkuvani tuon kesän ajan.

Cephin alustariippumattomuus ja ohjelmiston vapaus on syynä sille, miksi Ceph on mielenkiintoinen vaihtoehto ja pohja kaikelle datantallennukselle ja käsittelylle. Se voi teoriassa korvata kaikki alustariippuvaiset suurien yhtiöiden tarjoamat datantallennuspalvelut tulevaisuudessa. [1.]

Ceph on toimiessaan houkuttelevampi vaihtoehto, kun käytettävät palvelinkoneet voivat olla mitä tahansa ja tästä syystä yleensä halvempia, kuin suurilta palvelintarjoajilta. Cephin kanssa ei tarvitse lukittautua vain yhden yrityksen tuotteisiin. Kaikki tallennuskapasiteetti näkyy samanlaisena Vmwaren ja muiden pääkäyttäjien puolella, vaikka laitteet sen alla vaihtuisivatkin. [1.]

Tässä insinöörityössä raportoin käytännön selvitystä, jota suoritin CSC:n laitteilla, miten hyvin Ceph ja Vmware toimivat yhdessä ja onko tämä vaihtoehto myös käytännössä parempi ja jo nykypäivänä järkevä ratkaisu.

Opinnäytetyössä käyn läpi myös, miten Ceph voi korvata muita perinteisiä tallennusmenetelmiä ja miten se toimii esimerkiksi toisen virtuaalipalveluita tarjoavan OpenStackin pohjalla.

2 Ceph

2.1 Tausta

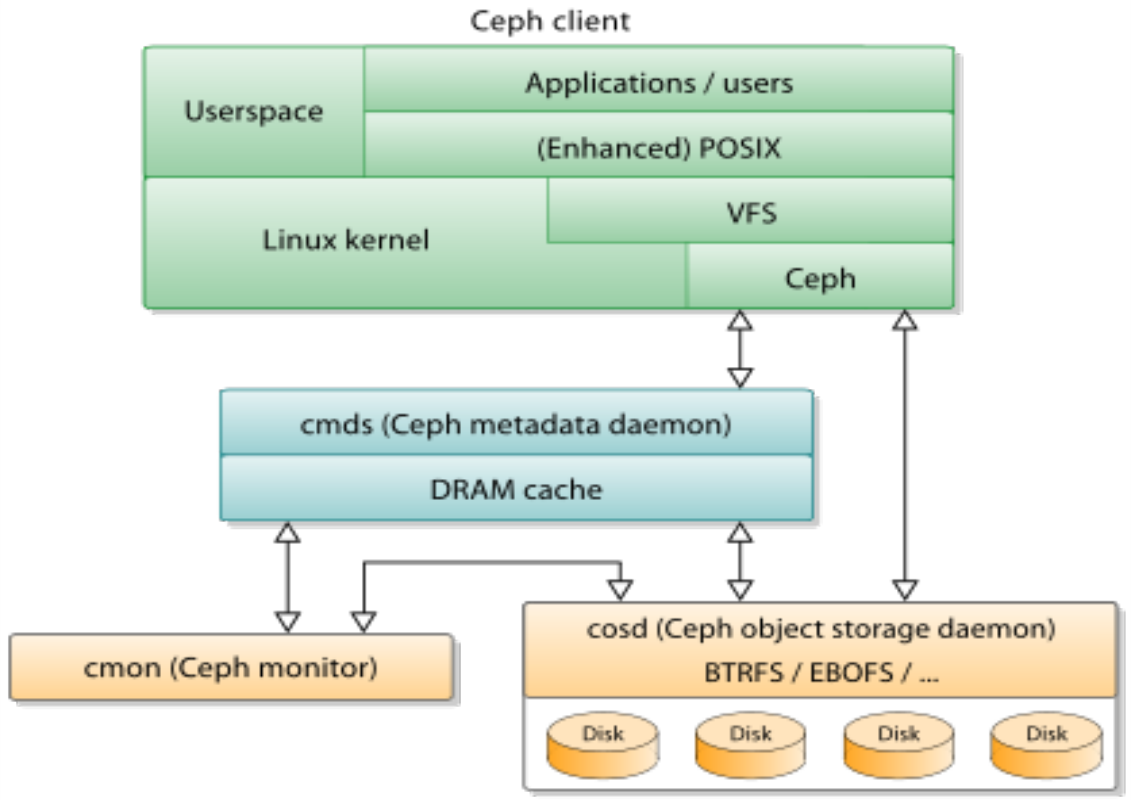
Ceph on yksinkertaisuudessaan vapaan lähdekoodin ohjelmistopohjainen tallennusjärjestelmä, joka tarjoaa helposti laajennettavaa tallennustilaa eri muodoissa. Cephiä voi

käyttää objekti-, lohko- ja tiedostopohjaisesti, eli englanniksi tunnetummin **object**-, **block**- ja **file storage** -muodoissa. Se on helposti laajennettavissa jopa eksatavun kokoluokkaan, eli tuttavallisemmin ~1000 petatavua, tai ~1000 000 teratavua. [2.; 3.; 4.]

Cephin suurin myyntivaltti on sen alustariippumattomuus ja hinta. Ceph-ympäristön voi pystyttää millä tahansa palvelintietokoneilla, joko isojen valmistajien palvelinkaapeilla (engl. rack) tai halvoilla yhden piirilevyn Raspberry Pi -koneilla. Nämä voivat toimia jopa yhdessä. Ceph vapaana ohjelmistona ei myöskään kustanna sen käyttäjälle mitään, toisin kuin kilpailijoiden vastaavat objektidatapohjaiset ratkaisut, kuten esimerkiksi Dellin Isilon-järjestelmä. [5.]

2.2 Cephin osat

Yksinkertainen Ceph-ympäristö voidaan pystyttää yhdelläkin tietokoneella, mutta kattava ja vikasietoinen ympäristö koostuu vähintään viidestä osasta, jotka on hyvä jakaa viidelle eri palvelinkoneelle. Näistä viidestä kolme on varsinaisia tallennusalustoja, joilla pyörii Cephin objektitallennuslaitteita (engl. object storage device) eli OSD:itä. Yleisenä ohjesääntönä nämä OSD:t vastaavat yleensä yhtä kiintolevyä palvelintietokoneessa, mutta voidaan asettaa vastaamaan myös useaa kiintolevyä tai esimerkiksi puolikasta kiintolevyosiota. Sen lisäksi ympäristössä tulee olla vähintään yksi Ceph-monitorilaite, joka seuraa tallennuskoneiden tilaa ja sen OSD:iden tilannetta. Se varmistaa esimerkiksi, että kaikki OSD:t ovat toiminnassa ja välittää tiedon vikatilanteista järjestelmälle niin, että järjestelmä voi kopioida epäkunnossa olevan kiintolevyn ja sitä vastaan OSD:n datan muualle turvaan. Viimeinen ja viides osa on metadata server daemon (MDS), eli metadatapalvelin, joka mahdollistaa tavallisten POSIX-järjestelmien komentojen, kuten **cp**- ja **mv**-käytön kuormittamatta itse tiedostopalvelimia. [3.]



Kuva 1. Havainnollistava kuva OSD-koneiden ja Monitor-koneiden suhteesta käyttäjään. [4; 6]

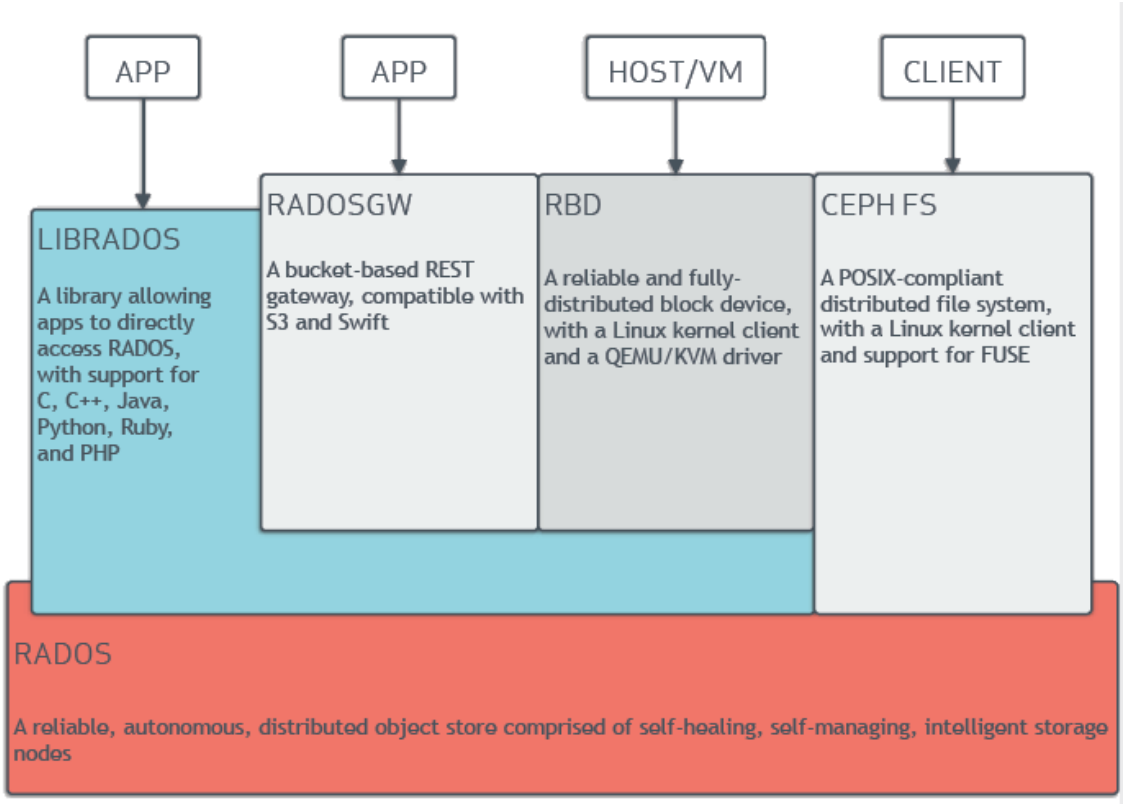
Ceph-ympäristöä laajennettaessa OSD-koneiden määrää voidaan kasvattaa ja niiden lisääntyessä on syytä lisätä myös monitor-koneita. Perusasetuksillaan Ceph kopioi dataa kolmelle koneelle. Tästä syystä pienessäkin ympäristössä OSD-koneita on syytä olla kolme. Jos koneet ovat erillään, ne ovat vikasietoisempia, kun yhden tai kahdenkin koneen hajotessa data pysyy vielä yhdellä koneella tallessa. Laajennettaessa on myös suositeltavaa kasvattaa monitor-koneiden määrä kolmeen. [7.]

2.3 Cephin toiminta

Ceph tallentaa kaiken datan objektimuodossa OSD:eille, ja sen ydin on RADOS (Reliable Autonomic Distributed Object Store). Tätä dataa voidaan käsitellä joko libradoksen läpi, tai CephFS:n läpi, joista jälkimmäinen tarjoaa suoraan tiedostopohjaisen käyttöliittymän käyttäjälle. Suuri Cephin etu on se, että näitä kaikkia ratkaisuja voidaan käyttää samanaikaisesti, koska kaikki tavat kääntyvät lopulta Radoksen tukemaksi ob-

jektidataksi. Se tekee Cephistä helposti laajennettavan moniin eri tallennustarkoituksiin ja tapoihin. [8.]

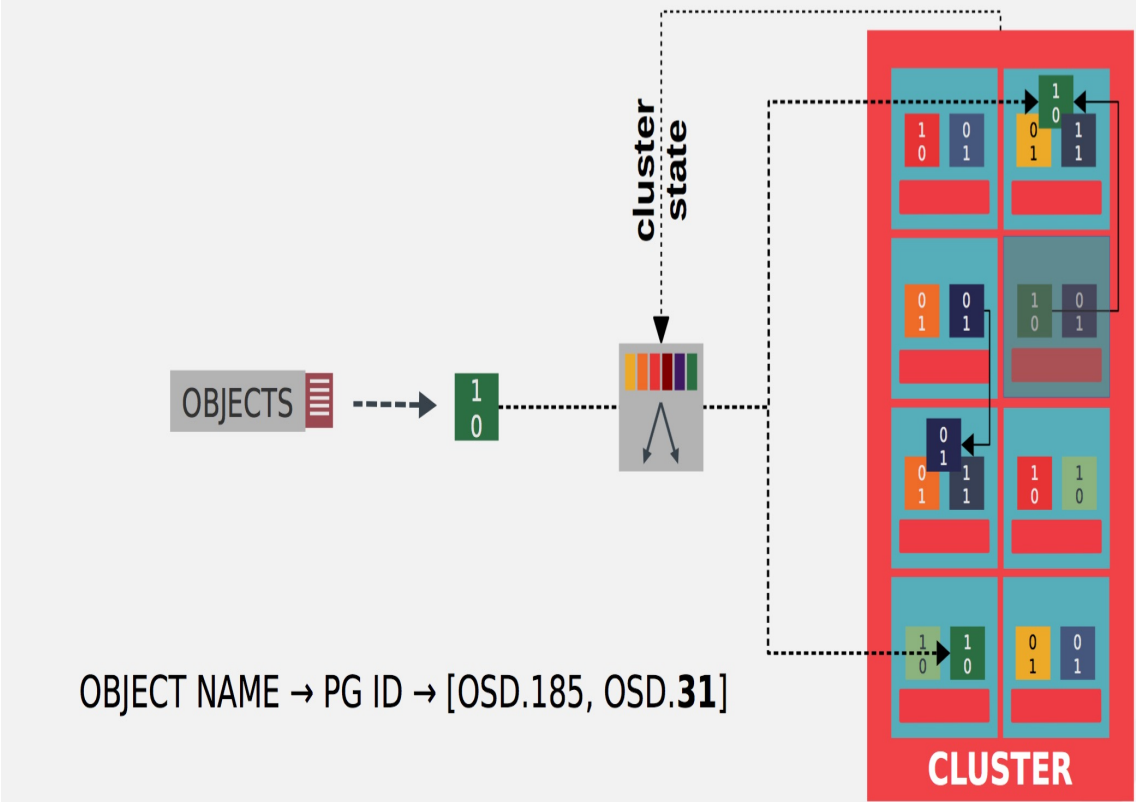
Librados mahdollistaa objektirajapinnan käytön suoraan, ja se tukee monia ohjelmointikieliä. Libradoksen avulla Cephin objektidataa voidaan käyttää myös Amazonin S3:lla tai Swiftillä REST API:n läpi. Kolmas vaihtoehto on hyödyntää Cephin objektidataa Rados Block Devicen läpi, jonka avulla Ceph näkyy ylläpitäjälle lohkotallennuslaitteena, jota juuri Vmware ja OpenStack hyödyntävät. OpenStack hyödyntää Linuxin omaa virtualisointiominaisuutta KVM:ää, jota RBD tukee suoraan. Kuva 2 havainnollistaa näiden osien roolia yhdessä. [8.]



Kuva 2. Ceph:n RADOS ja sen tavat keskustella eri väylien läpi. [20]

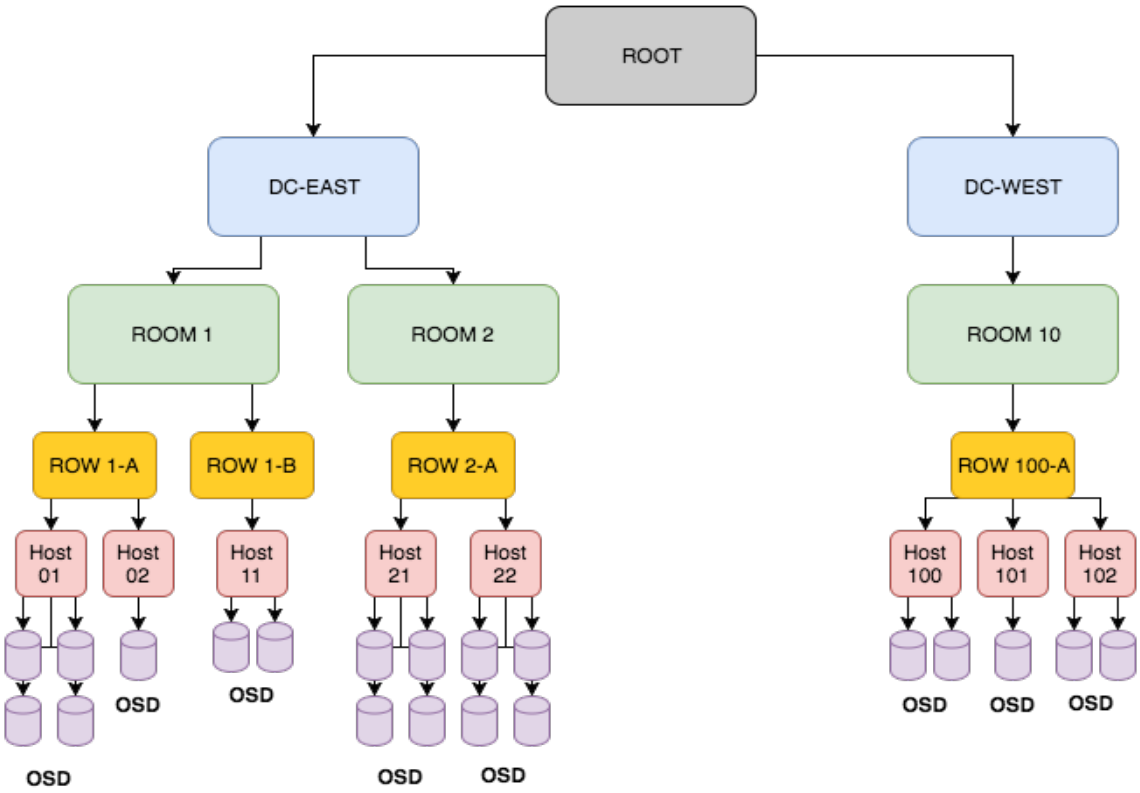
Cephin datan jakamiseen OSD:llä on kolme tärkeää komponenttia: **Pools**, **Placement Groups** (PG) ja **CRUSH Map**. Ceph tallentaa objektidatansa Pooleihin, loogisiin ryhmiin, jotka määrittävät Placement Groupien ja datakopioiden määrän ja CRUSH-säännön. Näistä pooleista voidaan ottaa myös varmuuskopioita. Placement Groupit asettavat objekteja ryhmiin ja helpottavat näin ollen poolien toimintaa. Pool yksin ei voi seura-

ta miljoonien yksittäisten objektien sijainteja ilman huomattavaa suorituskykyä ja rasitusta. CRUSH-mapit määrittävät, miten data jaetaan eri OSD:eiden kesken, jotta yksittäiset OSD:t eivät rasitu liikaa. [9.]



Kuva 3. Havainnollistava kuva siitä, miten CRUSH map siirtää dataa OSD:eiden välillä. [10]

Kuvassa 3 CRUSH-map säännöillä voidaan ohjata dataa niin, että kopiot datasta pysyvät aina eri palvelinhuoneessa tai eri palvelintornissa, jotta kaikki kolme kopiota datasta eivät ole alttiita samalle koneelle tai samalle huoneelle tapahtuvalle odottamattomalle vialle tai onnettomuudelle (esim. tulipalo). Kuvassa sininen ja vihreä objekti siirretään OSD:eiltä toisille CRUSH-mapissa asetetun säännön mukaan. Kuva 4 havainnollistaa, miltä suurempi Ceph-ympäristö voisi näyttää ja miten CRUSH map -säännöillä tätä dataa voidaan jakaa laajemmaltikin. [10.]



Kuva 4. Havainnollistava kuva Ceph-ympäristöstä suuressa yrityksessä, jossa on useampia datakeskuksia ja palvelinhuoneita. [10.]

Kuvan 4 tapauksessa datakeskukset DC-EAST ja DC-WEST voivat sijaita esimerkiksi eri kaupungeissa ja CRUSH map -säännöillä datakopiot voidaan jakaa näiden välille niin, että toisen datakeskuksen tuhoutuessa täysin kopioita on vielä olemassa. Kuvan 3 esimerkkiä käyttäen vihreä ja sininen objekti, jotka ovat yksittäisiä objekteja kolmesta kopiosta, voitaisiin siis siirtää DC-EAST-datakeskuksesta DC-WEST-datakeskukseen uudella CRUSH map -säännöllä. [10.]

Ceph on teoriassa luotettava pohja tallentaa dataa, joka tallentamis- ja käyttötapojen monipuolisuutensa ja sisäisen toimintansa vuoksi vaikuttaa erittäin varmalta ratkaisulta tulevaisuutta silmällä pitäen. [9]

3 VMware

3.1 Tausta

Vmware on yritys (VMWare Inc.) joka valmistaa monia pilvi-infrastruktuuriin ja verkkoyhteyksiin liittyviä palveluita. Näistä opinnäytetyötä varten keskeisin on Vmwaren KVM:ää vastaava Hypervisor ESX ja sen hallinnointiin käytettävät vSphere ja vCloud. [11.; 12.]

Virtualisointi on huomattavasti halvempaa kuin yksittäiset fyysiset palvelimet. Kun laitteet hajoavat, asiakas joutuu hankkimaan uusia osia, mikä aiheuttaa kustannuksia itse osista ja siihen liittyvästä ylläpitotyöstä. Kun asiakkaalle tarjotaan virtuaalipalvelimia, ne voivat sijaita samalla fyysisellä palvelimella ja ovat tarvittaessa lennosta siirrettäviä toiselle palvelimelle, jos alkuperäiseen palvelimeen tulee vika. Kun sama kone voi tarvittaessa palvella useampaa asiakasta, se on myös kustannustehokkaampaa. Monesti myös virtuaalikoneiden ylläpito on helpompaa, kun levytilan, prosessorien ja muistien muutokset asiakkaan päässä hoituvat suoraan ohjelmiston kautta, joka jakaa resursseja virtuaalikoneille. [13.]

Vmwaren ESX hyödyntää lohkotallennustilaa taustallaan ja yhteys tähän muodostuu joko iSCSI-protokollan tai Fibre Channel, eli FC-protokollan läpi. Loppukäyttäjän päässä tämä voidaan näyttää minä tahansa helpommin käytettävissä olevana tiedostojärjestelmänä, kuten esimerkiksi NFS tai NTFS. Vmwaren tapauksessa tämä tallennustila näytetään yleensä Vmwaren omana VMFS-tiedostojärjestelmänä. [14.]

Vmware tarjoaa lohkotallennuspalvelimien päälle oman ohjelmistopohjaisen virtuaalisen SAN-ympäristön (vSAN), joka mahdollistaa datan jakamisen helposti eri virtuaalikoneiden kesken. Se muuttaa fyysiset kiintolevyt virtuaaliseksi tallennusaltaaksi, jota voidaan vapaasti pilkkoa osiin ylläpidon ja asiakkaiden omien tarpeiden mukaan. Tämä ajaa osin samaa asiaa kuin Ceph, mutta pienemmällä latenssilla ja optimoidulla suorituskyvyllä. On siis jo alusta asti oletettavaa, että Ceph häviää Vmware-ympäristön suorituskyvylle, mutta tarjoaa silti ominaisuuksia, jotka voivat suorituskyvyn ollessa siedettävällä tasolla olla houkuttelevia yrityksille. [15.; 16.]

3.2 Vaihtoehdot tallennustilalle

Lohkotallennustila ostetaan yleensä suurilta palvelintarjoajilta, kuten Delliltä, HP:ltä, NetApp:ltä tai Huaweiltä. Tässä tapauksessa ostaja on kiinni valitsemansa tarjoajan palvelimissa ja lisensseissä, joihin yleensä sisältyy muiden muassa huoltotukisopimus. Tämä on helppo tapa saada huoletonta tallennustilaa, kun suuri yritys takaa sen toimivuuden sopimuksen ajan. [17.; 18.; 19.]

Moni näistä lohkotallennuslaitteita myyvistä suurista yrityksistä tarjoaa nykyään myös ohjelmistopohjaista objektitallennusratkaisuja, mutta toisin kuin Ceph, ne eivät ole vapaita ohjelmistoja, jotka mahdollistavat minkä vaan laitteiston käyttämisen ilman ylimääräisiä lisenssimaksuja. Yleensä nämä ratkaisut myydään pakettina yritysten omien laitteiden kanssa ja sopimukseen kuuluu laajennusmahdollisuudet nimenomaan heidän omilla laitteillaan. [20.; 21.]

Ceph on vapautensa vuoksi houkutteleva kohde tutkia, miten se toimii Vmwaren tallennusalustana. Näin laitteisto, jota Vmwaren kanssa käytetään, ei ole riippuvainen yritysten lisensseistä, ja se voi olla mitä vain ja mahdollistaa esimerkiksi vanhojen ja uusien laitteiden yhdistämisen saman ympäristön alle. Tämä tekee laitteistopuolesta paljon halvemman, kun investointeja suuriin kokonaisuuksiin kerralla ei tarvitse tehdä. Se myös helpottaa ylläpitopuolta, kun laitteisto on joustavaa, mutta tallennustila kuitenkin näkyy VMwaren puolella samanlaisena kuin mistä tahansa muualtakin saatu tallennustila. [1.]

Esimerkiksi Dellin kilpaileva ja Cephiä vastaava objektitallennusjärjestelmä on Isilon. Teoriassa se tukee samanlaisia ominaisuuksia kuin Ceph:kin, mutta on sopimuspuolensa myötä lähes verrannollinen myös tavanomaisiin lohkotallennusratkaisuihin. Laitteisto ja Isilon-ohjelmisto tilataan pakettina, johon sisältyvät hinnasta riippuen eritasoiset huoltosopimukset. Se ei tarjoa Cephin kaltaista joustavuutta, mutta voi silti olla houkuttelevampi vaihtoehto esimerkiksi juuri huoltotukensa ansiosta. [1.; 21.; 22.]

Cephissä on myös huonot puolensa. Pääasiassa Vmwaren ja Cephin välinen tuki ei ole erityisen hyvää, ja vaikka yhteys niiden välillä toimiikin, sen optimoimiseksi ei ole tehty juuri mitään, eikä sitä tueta virallisesti. Tässä syitä on monia, mutta kun käyttöjärjestelmäyritykset Red Hat ja SUSE ajavat Cephin kehitystä, ja Vmware itsekin on käyttöjär-

jestelmäpuolella kilpailijana, niin Vmwaren ja Cephin välisen suorituskyvyn optimointiin ei haluta käyttää resursseja. [1.]

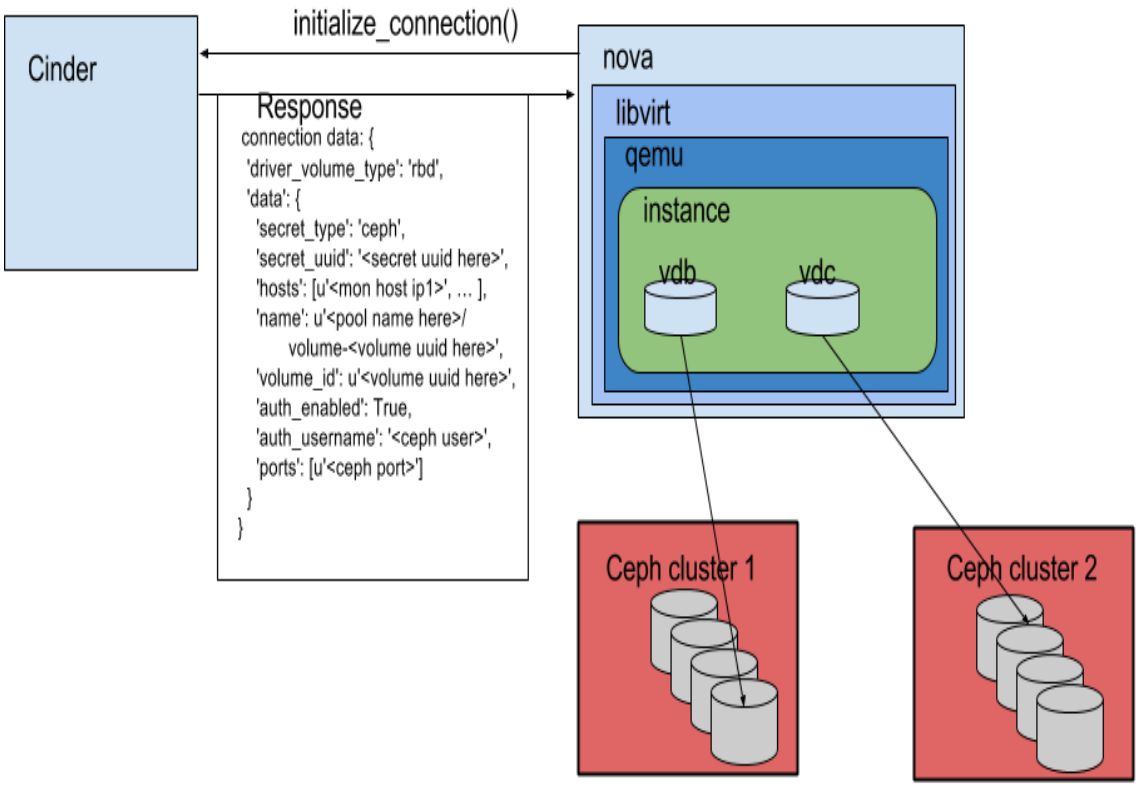
Cephiä käyttäessä myös levytilan käytössä ilmenee suurempaa latenssia eli viivettä, kuin perinteisillä SAN-ratkaisuilla, ja suurin osa Cephin halpuudesta tulee siitä, että siinä voi käyttää rajatta eri laitteita, usein vanhemmilla ja hitaammilla HDD-kiintolevyillä. Näiden suorituskyky uusiin SSD-kiintolevyihin verrattuna on huomattavasti huonompi, ja vaikka Ceph-ympäristön loisikin uusilla SSD-kiintolevyillä, se ei vieläkään poista Cephin ominaista latenssia. [1.]

4 Cephin käyttö

4.1 Cephin käyttö OpenStackin tallennusalustana

CSC käyttää Cephiä nykyisen OpenStack-ympäristömme tallennusalustana ja on tästä syystä ollut puheenaiheena sille, voisiko se toimia myös muiden palveluiden tallennuspohjana. OpenStack on Vmwarea vastaava virtualisointiympäristö, jossa käyttäjä tai ylläpitäjä voi luoda itselleen tai asiakkailleen virtuaalikoneita. OpenStack on yleensä käytetympi asiakkaille luotuna pilvipalveluna, kun taas Vmwarea pidetään enemmän yritysten sisäisenä virtualisointipalveluna. Suoria kilpailijoita nämä palvelut eivät ole, ja esimerkiksi Vmware on yksi suuria OpenStackin kehittäjiä. [23.]

Cinder on OpenStackin lohkotallennusta tarjoava palvelu, jonka pohjalla Cephiä voidaan käyttää tallennustilana. Nova on virtualisoinnista vastaava palvelu, joka lähettää pyynnöt Cinderille ja Cinder puolestaan pyynnön käsiteltyään takaisin Novan kautta Cephille. Tähän pyyntöön sisältyy esimerkiksi Poolin nimi ja järjestelmän IP-osoite. Kuva 5 havainnollistaa näiden yhteyttä Cephiin. [24.]



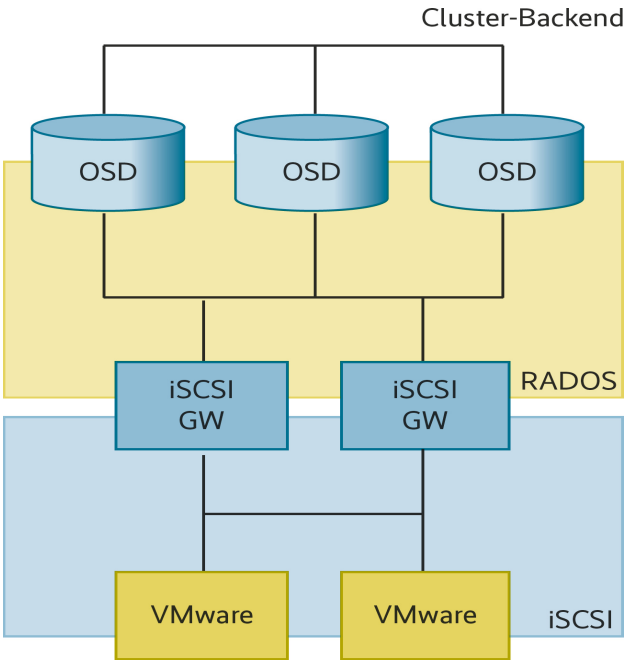
Kuva 5. Cephin käyttö OpenStackissa. [25.]

Ceph on suosittu pohja OpenStackille, ja Ceph itsekin ajaa itseään nimenomaan hyvänä vaihtoehtona OpenStackin tallennusalustaksi. Cephin ominaisuudet ja suorituskyky ovat hyvin sopivia esimerkiksi tietokantojen, median ja arkistojen ylläpitoon muun muassa juuri Cephin laajennettavuuden vuoksi. Kirjoitus- ja lukunopeudet ovat Cephin kanssa hyviä, ja HDD-kiintolevylaajennettavuutensa ansiosta tilaa on hintaan nähden paljon. [26.]

Ceph on OpenStackin alustana jo CSC:nkin käytössä hyväksi todettu ja Vmware ja tarkemmin ilmaistuna sen vSphere-ohjelmisto on OpenStackin kaltainen järjestelmä samankaltaisilla tarpeilla. CSC halusi tutkia, miten hyvin Ceph toimii muiden palvelujen tallennusalustana. Näistä ensisijaiseksi testikohteeksi valittiin juuri Vmware.

4.2 Cephin käyttö VMwaren tallennusalustana

Vmwaren ja OpenStackin tavat käyttää Cephiä ovat samankaltaiset. Tarkoituksena on tarjota Cephin lohkotallennustilaa suoraan RBD:n kautta Vmwarelle NFS-tiedostojärjestelmän lävitse ja sitten näyttää tämä tallennustila käyttäjälle XFS-tiedostojärjestelmänä, jossa voidaan suorittaa muiden muassa suorituskykymittauksia. Vmwaren ja Cephin välille ei ole mahdollista saada suoraa keskusteluyhteyttä, joten jokin keskusteluyhteyden mahdollistava tiedostojärjestelmä tai portti (engl. gateway) on pakollinen. Kuva 6 havainnollistaa tätä Cephin ja Vmwaren yhteyttä. [1.; 27.]



Kuva 6. Cephin käyttö Vmwaressa. [28]

Kuvassa 6 NFS-tiedostojärjestelmän sijaan se esittää Vmwaren ja RBD:n välillä käytettävän iSCSI-porttia, joka on yksi monista Cephin sisäänrakennetuista metodeista käyttää RBD:tä. Omassa ympäristössäni päädyin NFS-ratkaisuun monien Cephin sähköpostilistan käyttäjien suosituksista johtuen, jossa iSCSI-porttia pidettiin epävakaampana ratkaisuna. [29.; 30.]

Cephiä on käytetty onnistuneesti Vmwaren kanssa, mutta laajoja tuloksia siitä, kuinka hyvin se toimii, ei ole. Insinöörityöni tarkoitus on mitata tätä suorituskykyä ja arvioida,

onko Cephin kustannustehokkuus ja suorituskyky tarpeeksi hyvä verrattuna Vmwaren ja Hitachin kokonaisuuteen, että sillä voitaisiin tulevaisuudessa korvata jopa kaikki muu lohkotallennustila. [31.]

5 Käytännön selvitys

5.1 Tausta

Loin CSC:llä ympäristön, jossa on kaksi virtuaalikonetta. Toinen virtuaalikoneista käytti Hitachin SAN-lohkotallennustilaa, jonka päällä koko Vmware-ympäristömme pyörii, ja toinen Cephiä, joka luotiin eritoten tätä selvitystä varten.

Cephin puolella jouduin olosuhteiden asettamasta pakosta käyttämään kehitysympäristöämme, jossa ei ollut käytössä yhtään SSD-kiintolevyä. Tässä ympäristössä muutosloki (engl. Journal) oli samalla HDD-kiintolevyllä kuin datakin, ja tiesin sen vaikuttavan kielteisesti tuloksiin, mutta tasatakseni suorituskykymittauksia loin Hitachin levyjärjestelmän päälle tulevan Vmware-virtuaalikoneen myös HDD-kiintolevylle meidän niin sanottuun ”capacity”-ympäristöön. Tätä ympäristöä ei oltu missään vaiheessa tarkoitettu varsinaista suorituskykyä varten vaan suurien datamäärien tallentamista varten. Tästä huolimatta hypoteesi oli, että Vmware Hitachin levyjärjestelmällä tulee suoriutumaan suorituskykytesteistä Cephin levyjärjestelmää paremmin. [32.]

Jaoin Cephistä Rados Block Devicen Vmwaren käyttöön, ja tämä tila näytettiin virtuaalikoneilla NFS exportteina, jotka mountattiin XFS-tiedostojärjestelmänä suorituskykymittauksia varten. Suorituskykyä varten mittasin vain kirjoitus- ja lukunopeutta sekä latenssia. Suuremman luokan luotettavuusmittauksia ja vastaavia suurempaa aikajaksoa vaativia mittauksia ei tätä insinöörityötä varten tehty, mutta viittauksia Cephin luotettavuudesta ja toimivuudesta saatiin vuoden mittaan Cephin ylläpitopuolella.

Komennot, joilla tämä Rados Block Device luotiin ja jaettiin Vmwarelle on esitetty alla kohdassa Esimerkkikoodi 1. Rivit ovat kommentoituja ja selittävät, mitä komennot tekevät.

```
# Luodaan uusi Rados Block Device
rbd create lohko --size 4096 --image-feature layering [-m {mon-IP}] [-k
/path/to/ceph.client.admin.keyring]
rbd map lohko --name client.admin [-m {mon-IP}] [-k /path/to/ceph.client.ad-
min.keyring]

# Käytetään XFS-tiedostojärjestelmää luodun RBD:n käyttämiseen.
mkfs.xfs -m0 /dev/rbd/rbd/lohko

# Luodaan hakemisto, joka jaetaan myöhemmin Vmwarelle.
mkdir /mnt/ceph-rbd
mount /dev/rbd/rbd/lohko /mnt/ceph-rbd

# Tämä hakemisto näytetään Vmwarelle NFS exporttina "/etc/exports"
# tiedostossa.
/mnt/ceph-rbd {ip}(rw,sync,no_root_squash,no_subtree_check)
```

Esimerkkikoodi 1. Ceph Rados Block Devicen luominen.

Tämä luotu ”NFS export” voidaan valita Vmwaren puolella uutena ”datastorena” mittauksessa käytettävälle virtuaalikoneelle, ja täten virtuaalikone käyttää tallennuspohjanaan Cephiä.

Toinen mittauksessa käytetty Vmware-virtuaalikone, joka käytti Hitachin SAN-tallennustilaa, luotiin yksinkertaisesti vSpheren käyttöliittymästä, jossa yllä mainittu datastore oli valmiiksi vSphereen asetettu Hitachin levytila. Tälle vaihtoehtoina oli SSD-kiintolevyillä toimiva ”performance” ja HDD-kiintolevyillä toimiva ”capacity”, joista valitsin jälkimmäisen. Tiedostojärjestelmänä käytettiin Cephin tavoin XFS:ää, jotta sen puolesta ei muodostuisi eroja ympäristöjen välille.

5.2 Suorituskykymittaukset

Käytin suorituskykymittauksissa sitä varten tarkoitettuja ja hyödyllisiä työkaluja fio ja dd. Fio eli ”Flexible Input/Output Tester” mahdollistaa esimerkiksi läpisyötön mittaamisen, eli kuinka monta megatavua sekunnissa virtuaalikoneella voidaan kirjoittaa dataa sen taustalla olevalle tiedostopalvelimelle. Tätä voidaan mitata lähettämällä tiedostopalvelimelle suhteellisen suurikokoinen tiedosto, jossa on suurikokoisia lohkoja. Tämä voisi vastata esimerkiksi suurta mediatiedostoa. [33.]

Toisena mittauskohteena mittasin siirräntää (engl. input/output), jossa mitataan, kuinka monesti tieto kulkee järjestelmään ja sieltä ulos sekunnissa. Tätä voidaan mitata lähettämällä suuri määrä pienilohkoisia tiedostoja, jotka vastaisivat todellisuudessa esimerkiksi logi- ja dokumenttitiedostoja. [33.]

Toisena työkaluna käytin Unix-pohjaista dd-työkalua, jolla mittasin läpisyöttöä ja latenssia Hitachilla ja Cephillä. Sen pääasiallinen tarkoitus on kopioida tiedostoja paikasta A paikkaan B ja tässä suorituskykymittauksessa siirrän /dev/zero-”tiedostolla” nollia eri lohkokoilla. Iso lohko mittaa tässäkin tapauksessa isoa tiedostoa, mutta pienellä lohkolla testattessa tarkoituksena on saada tieto siitä, kuinka nopeasti pieni tiedosto kulkee paikasta toiseen, eli kuinka pitkä latenssi tai viive virtuaalikoneen ja tiedostopalvelimen välillä on. [34.]

FIO-ohjelmalla suoritetut mittaukset tehtiin seuraavilla kohdassa Esimerkkikoodi 2 näkyvillä komennoilla. Tiedostojen koot ovat suuria, neljä ja kymmenen gigatavua, mutta yksittäisen lohkon koko vaihtuu riippuen mittauksesta. Siirräntää mitattaessa lohkon koko on 4 kilotavua ja läpisyöttöä mitattaessa 4 megatavua.

```
# Peräkkäinen levylle kirjoitus IOP/s
fio --randrepeat=1 --ioengine=libaio --direct=1 --gtod_reduce=1 --name=test
--filename=test --bs=4k --iodepth=256 --size=4G --readwrite=write –ramp_time=4

# Peräkkäinen levyltä luku IOP/s
fio --randrepeat=1 --ioengine=libaio --direct=1 --gtod_reduce=1 --name=test
--filename=test --bs=4k --iodepth=256 --size=4G --readwrite=read –ramp_time=4

# Satunnainen levylle kirjoitus IOP/s
fio --randrepeat=1 --ioengine=libaio --direct=1 --gtod_reduce=1 --name=test
--filename=test --bs=4k --iodepth=256 --size=4G --readwrite=randwrite –
ramp_time=4

# Satunnainen levyltä luku IOP/s
fio --randrepeat=1 --ioengine=libaio --direct=1 --gtod_reduce=1 --name=test
--filename=test --bs=4k --iodepth=256 --size=4G --readwrite=randread –
ramp_time=4

# Peräkkäinen levylle kirjoitus MB/s
fio --randrepeat=1 --ioengine=libaio --direct=1 --gtod_reduce=1 --name=test
--filename=test --bs=4M --iodepth=256 --size=10G --readwrite=write –
ramp_time=4

# Peräkkäinen levyltä luku MB/s
fio --randrepeat=1 --ioengine=libaio --direct=1 --gtod_reduce=1 --name=test
--filename=test --bs=4M --iodepth=256 --size=10G --readwrite=read –ramp_time=4

# Satunnainen levylle kirjoitus MB/s
fio --randrepeat=1 --ioengine=libaio --direct=1 --gtod_reduce=1 --name=test
--filename=test --bs=4M --iodepth=256 --size=10G --readwrite=randwrite –
ramp_time=4

# Satunnainen levyltä luku MB/s
fio --randrepeat=1 --ioengine=libaio --direct=1 --gtod_reduce=1 --name=test
--filename=test --bs=4M --iodepth=256 --size=10G --readwrite=randread
--ramp_time=4
```

Esimerkkikoodi 2. FIO-komennot, joilla suorituskykyä mitattiin. Komentojen erot on merkitty kommenttiriveillä komennon yläpuolella.

Mittauksissa vertaillaan tilannetta, jossa lohkot ovat perätysten ja satunnaisesti hajautettu. Satunnaisesti hajautetuissa lohkoissa tallennusalustalla on suurempi työ löytää näitä lopullisen tiedoston palasia, ja satunnainen haku on täten lähes poikkeuksetta hitaampi kuin peräkkäinen haku. Peräkkäisessä haussa lohkojen etsimiseen ei kulu ylimääräistä aikaa. [35.]

Komennoissa on paljon osia, joilla voi olla vaikutusta tuloksiin. Muiden muassa ”iodepth”-parametri määrää, kuinka monta input/output-käskyä fio lähettää kerrallaan Esimerkkikoodi 2:n komennoissa määriteltyyn testitiedostoon. Tämän kasvattamisesta ei välttämättä ole hyötyä hitaammilla HDD-kiintolevyillä, mutta SSD-kiintolevyillä on suorituskykyä käsitellä satoja kerralla. ”direct”-parametri määrää sen, käytetäänkö suoraa käyttöjärjestelmän kirjoitustapaa vai esipuskuroitua kirjoitustapaa. Esipuskuroimaton kirjoitustapa on lähes poikkeuksetta parempi suorituskykymittauksia varten, sillä kirjoitus tapahtuu suoraan ilman ylimääräistä puskurointioperaatiota, mutta kaikki käyttöjärjestelmät eivät tue sitä. ”gtod_reduce”-parametri vähentää tämän hetkisen kellonajan selvittämiseen käytettyjä kutsuja, jotka voivat helpottaa komennon tuottamaa rasitusta ja antaa todenmukaisemman suorituskykyarvion. ”ioengine”-parametrilla voimme osoittaa komennon käyttämään Linux-käyttöjärjestelmän omia kirjoitusmetodeja. Tämä vaatii toimiakseen myös direct-parametria. [33.; 36.]

Päädyin näissä komennoissa lähteiden esimerkkeihin ja suosituksiin siitä, mitkä tuottavat luotettavimmat ja vertailukelpoisimmat tulokset. Mittaustulokset näkyvät taulukossa 1. Kaikki mittaustulokset ovat kolmen mittauskerran keskiarvoja, jotta tulokset olisivat luotettavampia. [33.; 36.]

Taulukko 1. Mittaustulokset siirrännälle. Työkalu: fio.

Luku/kirjoitustapa	**Vmware + Hitachi SAN Block Storage**	**Vmware + Ceph RBD Block Storage**
Peräkkäinen levylle kirjoitus	6978 IOP/s	993 IOP/s
Peräkkäinen levyltä luku	49552 IOP/s	33984 IOP/s
Satunnainen levylle kirjoitus	3632 IOP/s	227 IOP/s
Satunnainen levyltä luku	20002 IOP/s	12529 IOP/s

Mittauksissa huomataan heti suuria eroa jo käytössä olevan Hitachin levyjärjestelmän hyväksi. Cephin kanssa tulosten, varsinkin satunnaisen levylle kirjoituksen, odotettiinkin olevan alhaisempia, sillä Ceph kirjoittaa dataa kolmelle eri OSD:lle ja kun kirjoitettavia lohkoja on paljon, se rasittaa järjestelmää enemmän. Teoriassa läpisyötön yhteydessä erojen ei pitäisi olla yhtä mittavia, koska samanlaista rasitusta ei muodostu, kun lohkot ovat isoja ja niitä on vähemmän. Seuraavaksi mittasin näiden eri kirjoitus- ja lukutapojen läpisyöttöä taulukossa 2.

Taulukko 2. Mittaustulokset läpisyötölle. Työkalu: fio.

Luku/kirjoitustapa	**Vmware + Hitachi SAN Block Storage**	**Vmware + Ceph RBD Block Storage**
Peräkkäinen levylle kirjoitus	92 MB/s	38 MB/s
Peräkkäinen levyltä luku	552 MB/s	578 MB/s
Satunnainen levylle kirjoitus	82 MB/s	35 MB/s
Satunnainen levyltä luku	646 MB/s	403 MB/s

Tulokset vahvistavat aiempaa teoriaa. Läpisyötössä Ceph ei ole läheskään yhtä moninkertaisesti Hitachia jäljessä, ja peräkkäisen levyltä lukemisen kohdalla Cephin keskiar-

vo nousi jopa Hitachin yli. Tämä oli mielenkiintoinen havainto, sillä Cephin uuden version parantaessa kirjoitusnopeutta Cephin hyvät puolet halpuutensa ja laajennettavuutensa ansiosta voi hyvinkin nopeasti muodostua suurien yritysten levyjärjestelmiä paremmaksi vaihtoehdoksi.

Jatkoin testaamista vielä dd-työkalulla, jotta työkalusta ja sen asetuksista johtuvat virheet voidaan karsia parhaan mukaan pois. Tätä varten käytetyt komennot näkyvät esimerkkikoodissa 3.

```
# Peräkkäinen levylle kirjoitus MB/s
dd if=/dev/zero of=/mnt/testdrive/test1.img bs=1G count=1 oflag=dsync

# Peräkkäinen levylle kirjoitus, latenssi.
dd if=/dev/zero of=/mnt/testdrive/test2.img bs=512 count=1000 oflag=dsync
```

Esimerkkikoodi 3. dd-komennot, joilla suorituskykyä mitattiin.

Läpisyöttöä mitattaessa lohkon koko on valtava: yksi gigatavu, mutta lohkojen määrä on yksi. Latenssia mitattaessa lohkon koko on 512 kilotavua, mutta määrä on 1000. Tämän komennon lopulliset tulokset näkyvät taulukossa 3.

Taulukko 3. Suorituskyky mittaustulokset. Työkalu: dd.

Luku/kirjoitustapa	**Vmware + Hitachi SAN Block Storage**	**Vmware + Ceph RBD Block Storage**
Peräkkäinen levylle kirjoitus, läpisyöttö	165 MB/s	22.6 MB/s
Peräkkäinen levylle kirjoitus, latenssi	2.6 s	28.5 s

Nämä tulokset sivuavat fio-työkalulla todettuja suorituskykyetuja Hitachin levyjärjestelmän hyväksi. Läpisyötössä ero on hieman suurempi kuin fiolla mitattuna, mutta ei niin moninkertainen kuin siirräntää mitattaessa.

5.3 Tulosten analysointi

Tulosten perusteella jo huomataan, että Ceph ei suoriutunut lähes millään osa-alueella Hitachin levyjärjestelmää paremmin, mutta mittauksia tehdessä positiivista oli kuitenkin

huomata, että Ceph toimi Vmware-virtuaalikoneiden alustana muutoin moitteitta. Ilman vertailuarvoja virtuaalikonetta käytettäessä ei välttämättä edes tunnista, että mikään olisi muuttunut.

Lähempää tarkasteltuna aivan ongelmitta Cephin puolella ei kuitenkaan selvitty. Suorituskykymittauksia tehdessä ympäristön kanssa oli ongelmia, joihin emme ryhmänäkään löytäneet heti syitä tai ratkaisua. Ceph-ympäristössä kirjoitusnopeus saattoi aika-ajoin pysähtyä kokonaan, joka johti todella paljon alhaisempiin tuloksiin, kuin mitä keskiarvo muuten antoi odottaa. Syyksi tälle arvioimme jonkinlaista ruuhkanesto-ominaisuutta (engl. congestion), jossa järjestelmän vuon hallinta (engl. flow control) estää liian monen kyselyn läpi tulemisen. Tätä ei kuitenkaan tutkittu syvemmin ja harvinaisuutensa vuoksi sivuutimme sen. [37.]

Cephillä kirjoitusnopeudet saattoivat olla jopa 16 kertaa hitaampia kuin Hitachin levyjärjestelmällä. Tälle yksi suuri syy on heti todennäköisimmin se, että käytössäni oli vain CSC:n Ceph-testiympäristö, jolla ei ollut yhtään SSD-kiintolevyä. Vaikka myös Hitachin levyjärjestelmää käyttäessä käytin vain HDD-kiintolevyjä, Ceph-ympäristön suorituskykyä helpottaa huomattavasti se, jos kirjoittamattoman datan muutosloki, eli ”journal”, voidaan pitää SSD-kiintolevyllä, vaikka itse data kirjoitetaankin HDD-kiintolevylle. Varsinkin pienien lohkokokojen kirjoituksissa Journal avustaa datan kirjoitusta levylle paljon. [32.]

Vmware-ympäristömme varsinkin SSD-kiintolevyillä on äärimmäisen nopea ja viiveet tiedostojärjestelmän ja Vmwaren välillä ovat millisekuntien luokkaa. On siis syytä olettaa, että Ceph-ympäristön optimoiminen ei siltikään yltäisi Vmware-ympäristömme suorituskykyyn juuri Vmwaren tallennusalustana.

6 Cephin tulevaisuus CSC:llä

6.1 Cephin ongelmat

Käyttäessämme Cephiä sen myyntivalttina toimiva luotettavuus oli välillä kyseenalaista. Asiakkaan päässä näkyi aika-ajoin ongelmia, jossa heidän järjestelmänsä pysähtyivät, koska he eivät voineet kirjoittaa Ceph-järjestelmään. Tämä muistuttaa myös suorituskykymittauksissa tulleita kirjoituksen pysähdyksiä, mutta keksimämme ratkaisu ongel-

maan toimi asiakkaalle, ei suorituskykymittauksiin. Vaikka ongelman ydin on meille tuntematon, huomasimme, että jotkin OSD:t ottivat suuremman osan kuormasta kuin muut ja ajan kanssa niiden vastaavat kiintolevyt täyttyivät kauan ennen muita. Löysimme CERN:in Ceph-ympäristöä varten luodun skriptin heidän GitHub-sivultaan, jolla näitä OSD:n täyttymisen epätasaisuuksia voidaan tasoittaa. Skripti varmistaa, että mikään OSD ei ota leijonaosaa saapuvasta datasta. [38.]

Toinen ongelma oli palvelinympäristön mahdolliset toimintahäiriöt, kuten internetyhteyden katkeaminen, jolloin Ceph alkaa siirtää suuria määriä OSD:iden sisältämää dataa muualle. Tämä aiheutti suurta rasitusta järjestelmässä, ja se näkyi hitautena asiakkaan päässä. Ominaisuudella on myös puolensa, sillä se varmistaa, että kaikesta datasta on koko ajan kolme kopiota. Tämä voi kuitenkin johtaa myös tilanteisiin, jossa Ceph ei pysty käsittelemään samanaikaisesti suurta käyttäjän tuottamaa kuormaa. [39.; 40.]

6.2 Ceph Luminous

Kaikki suorituskykymittaukset insinöörityötä varten tehtiin kesän 2017 aikana, jolloin Cephin uusin versio oli vielä Jewel. Uusi versio Luminous julkaistiin elokuussa 2017 ja tuli CSC:lle käyttöön kehitysympäristössä noin kuukautta myöhemmin. Tämän uuden version tärkein ominaisuus oli kokonaan uusi objektitallennusmetodi BlueStore. BlueStore on kirjoitusnopeudeltaan kaksi kertaa nopeampi ja tarjoaa väitetysti muillakin osa-alueilla kaksinkertaisia parannuksia nopeudessa. [41.]

Suoritin suorituskykymittauksia eri ympäristössä, jossa kokoonpanoon ei enää kuulunut Vmware-virtuaalikonetta verratakseni BlueStoren tuomaa suorituskykyetua. Tämän mittauksen tulokset näkyvät taulukossa 4. Suoritin mittaukset fio-työkalulla ja täysin samoilla komennoilla kuin Vmware-ympäristöä mitattaessa, eli Esimerkkikoodi 2:n komennoilla. Ympäristö oli kehitysympäristö eri verkossa kuin Vmwaren mittauksissa. Tulokset Jewelin ja Luminouksen välillä ovat kuitenkin keskenään vertailukelpoisia.

Taulukko 4. Mittaustulokset siirräntä. Työkalu: fio.

Luku/kirjoitustapa	**Ceph Jewel ilman BlueStorea**	**Ceph Luminous BlueStoren kanssa**
Peräkkäinen levylle kirjoitus	306 IOP/s	514 IOP/s
Peräkkäinen levyltä luku	5881 IOP/s	18623 IOP/s
Satunnainen levylle kirjoitus	1982 IOP/s	15327 IOP/s
Satunnainen levyltä luku	7569 IOP/s	26442 IOP/s

Pienen lohkokoon tiedostoissa parannus aikaisempaan on huomattava ja sivuaa aikaisemmin todettua odotusarvoa kaksinkertaisista parannuksista, varsinkin satunnaisen kirjoituksen yhteydessä. Huomattava ja yllättävä huomio oli se, että ongelma kirjoituksen satunnaisen pysähtymisen kanssa oli hävinnyt Luminouksen puolella. Suoraa syytä tälle en löytänyt, vaikka kokeilin mittauksessa ja ympäristössämme erilaisia asetuksia. Lopullinen syy voi liittyä versiopäivitykseen tai itse BlueStore-teknologiaan, mutta se voi johtua myös uuden ympäristön pystytykseen liittyneistä muutoksista. Samaa ongelmaa ei myöskään ilmennyt läpisyöttöä mitattaessa, jonka tulokset näkyvät taulukossa 5.

Taulukko 5. Mittaustulokset läpisyöttö. Työkalu: fio.

Luku/kirjoitustapa	**Ceph Jewel ilman BlueStorea**	**Ceph Luminous BlueStoren kanssa**
Peräkkäinen levylle kirjoitus	257 MB/s	945 MB/s
Peräkkäinen levyltä luku	1021 MB/s	1780 MB/s
Satunnainen levylle kirjoitus	221 MB/s	987 MB/s
Satunnainen levyltä luku	1068 MB/s	1796 MB/s

Sama odotusarvo tulosten paranemisesta täyttyy myös läpisyötön kanssa. Tulokset peräkkäisen ja satunnaisen kirjoituksen välillä eivät ole suuria, mutta tämä on jossain määrin odotettavaa, koska lohkokoko on niin suuri, ettei sen satunnainen sijoittaminen tai lukeminen tuota paljoa ylimääräistä työtä.

6.3 Ceph muiden palveluiden korvaajana

Vaikka BlueStorea en päässyt suoraan mittaamaan Vmwarea vastaan, on sen aikaisempaan versioon verrattavat tulokset kuitenkin huomattavasti parempia. Tämä on suuri syy sille, miksi Ceph on jatkanut kasvuaan CSC:llä ja on merkittävässä osassa sen tulevaisuutta.

Ceph voisi jo nyt teoriassa korvata suurilta yritykseltä ostamamme levyjärjestelmät, ja voisimme tarjota vanhallakin laitteistolla yhtä paljon, ellei jopa enemmän lohkotallennustilaa. Uuden BlueStoren ja sen ylläpidossa huomaamamme vakauden vuoksi se on keskusteluissa nyt ja tulevaisuudessa. Cephin avulla voisimme edelleen käyttää vanhaa kalustoamme, hyödyntää sitä vapaasti ja emmekä olisi lukittautuneita sopimuksiin tiettyjen yritysten tuotteiden kanssa. Tämä tekee Cephistä houkuttelevan vaihtoehdon, kun Ceph itsessään ei tuota ylimääräisiä kustannuksia.

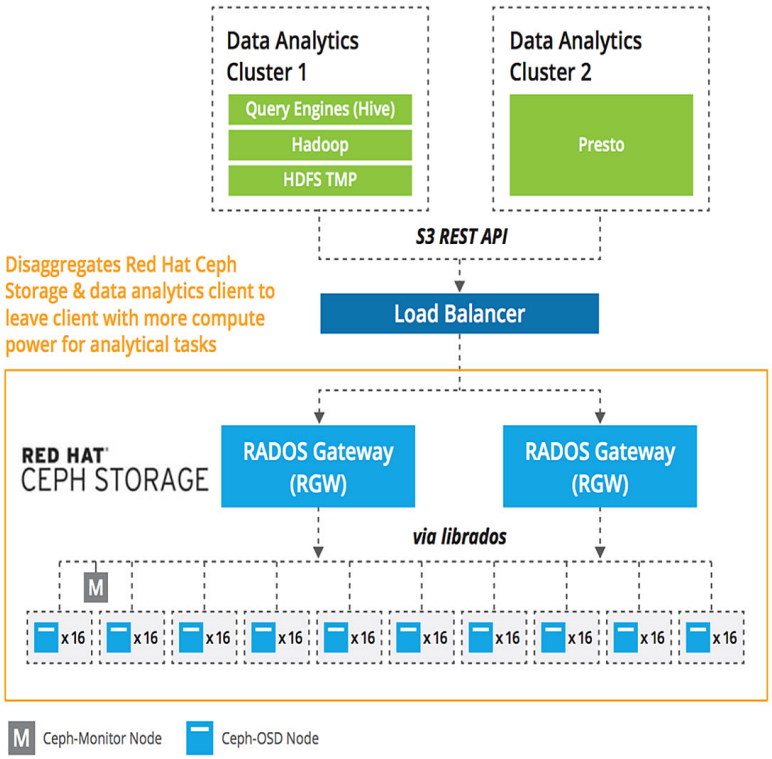
6.4 Ceph Suomen uuden supertietokoneen tallennusalustana

Suurin merkki Cephin läpimurrosta CSC:llä on sen valituksi tuleminen Suomen uuden 37 miljoonan euron investoinnin, supertietokoneen, tallennusalustana. Ceph on toiminut CSC:llä usean vuoden ajan OpenStackin tallennusalustana ja osoittautunut uusien versioidensa myötä suorituskykyiseksi, edulliseksi ja luotettavaksi.

CSC:n ollessa julkinen ja voittoa tavoittelematon yritys, on Ceph-teknologian avoimuus sen tuottamalle teknologialle sopivaa, ettei se ole lukossa tietyn palvelintarjoajan teknologioihin ja siihen liittyviin maksuihin. Supertietokoneen hankintaprojektin päällikkö Sebastian von Alfthan CSC:ltä kommentoi hankintaa ja osin Cephiä näin: [42.]

”Olemme innoissamme voidessamme tuoda uusimmat prosessori- ja kytkentäverkkoteknologiat hyödyttämään suomalaista tutkimusta yhdessä monipuolisen datanhallintaympäristön kanssa, joka perustuu avoimen lähdekoodin teknologioihin” [42]

Ceph tulee muodostamaan uudessa supertietokoneessa niin sanotun ”datalaken” tallennuspohjan. Tämä ”allas” tarjoaa asiakkaille, kuten esimerkiksi yliopiston tutkijoille helposti laajennettavan ja datantallennustarpeisiin sovellettavan pohjan, jossa tallennustilaa tutkimuksille voidaan tarjota jopa kymmeniä petatavuja. Tätä tallennustilaa voidaan helposti käyttää erilaisten rajapintojen läpi ja hyödyntää esimerkiksi ohjelmoidessa fysiikanmallinnus- ja kemiantutkimuksiin liittyviä sovelluksia, jossa käsiteltävää dataa on paljon. Kuva 7 havainnollistaa tätä käytännössä. [43.; 44.]



Kuva 7. Oranssilla on rajattu ”allas” ja sen yllä ovat sitä hyödyntäviä sovelluksia. [45]

Eritoten ”Big Datan” eli niin sanotun massadatan käsittely, jossa jatkuvasti kerääntyvää ja järjestelemätöntä dataa analysoidaan ja tutkitaan kasvattaa merkitystään ja sen pro-

sessointiin, eli datanlouhintaan ja muuhun siihen liittyvään tutkimustyöhön Cephin datalake tarjoaa täydelliset puitteet tallennusalustana. [46]

6.5 Loppusanat

Ceph oli kesällä 2017 omien suorituskykymittaukseni aikaan vielä jokseenkin epävakaa ylläpitäessämme sitä, ja mittauksissa huomattiin jonkinasteisia puutteita muita levyjärjestelmiä vastaan. Se on kuitenkin kehittynyt lyhyessä ajassa paljon ja sen suorituskyky ja luotettavuus ovat kasvaneet sen mukana. Tämä on tehnyt Cephin pohjimmaisesta tarkoituksesta olla maksuton, avoin ja tehokas ratkaisu yritykselle oikeasti varteenotettavan vaihtoehdon.

Opinnäytetyön ydinkysymykseen, toimiiko Ceph Vmwaren tallennusalustana, saatiin hieman kaksijakoinen vastaus. Se toimii, mutta kuten mittaukset osoittivat, ei kovin hyvin. Uudesta versiopäivityksestä ja BlueStore-teknologiasta huolimatta Vmware ei itsessään tue toimimistaan Cephin kanssa, ja jos tähän ei tule muutosta tulevaisuudessa, se tuskin koskaan tulee olemaan nopeampi ja varmempi vaihtoehto Vmwaren tarjoamille ratkaisuille. Suuri kysymys varsinkin suuryrityksille on myös Vmwaren tarjoama tuki siitä, että tuote varmasti toimii. Tämä on vain heidän tarjoamilleen ratkaisuille, joihin Ceph ei sisälly. Vmwaren vSAN:in ajaessa jo osittain samaa teknistä toteutusta kuin Ceph:kin, on Ceph tässä suljetussa ympäristössä omine hyvine puolineen kuitenkin liian rajallinen, eivätkä sen hyvät puolet pääse samalla tavalla esiin kuin muissa vapaammissa ympäristöissä. [16.]

Ceph on vapaa tallennusjärjestelmä ja sopii täten paremmin esimerkiksi vapaan virtualisointiympäristö OpenStackin rinnalle. Pienenkään yrityksen ei kannata sijoittaa Vmwareen vain jättääkseen heidän tarjoamansa tukipaketit pois ja käyttääkseen Cephiä. Pienemmälle yritykselle etenkin, jos virtualisointipuolella halutaan säästää, on loogisempaa säästää koko virtualisointiympäristön kanssa, ei pelkästään tallennusalustan kanssa. OpenStackin kaltaisessa ympäristössä Ceph on ominaisuuksiensa ja varsinkin uusien päivitystensä myötä osoittautunut erinomaiseksi ratkaisuksi. Cephiä on mahdollista käyttää myös muille lohko- tai objektitallennustilaa hyödyntäville palveluille, ja sen käyttö uuden superkoneen tallennusalustana todistaa sen käyttökelpoisuutta. Voimme tarjota asiakkaille esimerkiksi Amazon S3 -yhteensopivan webkäyttöliittymän, jota he voivat käyttää datan tallennukseen ja Ceph toimii tämän kanssa erinomaisesti. [47.]

Ceph tekee läpimurtoaan CSC:llä ja on vankkana osana sen tulevaisuutta datantallennuspuolella, joka ohjautuu objektipohjaisen tallennustavan puolelle. Ceph on joustava tapa yhdistää yrityksen sisäisiä tallennuspalveluja ja ulkoisia pilvipalveluita saman tehokkaan katon alle. Vmware jäänee kuitenkin omaksi kokonaisuudekseen. [48.]

Lähteet

1 Margaret Fisk. Ceph Storage for VMware vSphere. <http://virtunetsystems.com/ceph-storage-for-vmware-vsphere/> 14.09.2018. Luettu: 10.04.2019.

2 Ceph <https://searchstorage.techtarget.com/definition/Ceph> 07.2016. Luettu: 20.01.2019.

3 Ceph Storage Introduction. <https://ceph.com/geen-categorie/ceph-storage-introduction/> 05.12.2013. Luettu: 30.01.2019.

4 M. Tim Jones. Ceph: A Linux petabyte-scale distributed file system <https://www.ibm.com/developerworks/library/l-ceph/l-ceph-pdf.pdf> 04.06.2010. Luettu: 18.04.2019.

5 My Ceph Test Cluster Based on Raspberry Pi's and HP MicroServers. <https://louwrentius.com/my-ceph-test-cluster-based-on-raspberry-pis-and-hp-microservers.html> 27.01.2019. Luettu: 14.02.2019.

6 Kuvalähde. Yhteydessä lähteeseen #4. Teoriaa havainnollistava kuva. <https://en.wikipedia.org/wiki/Ceph_(software)#/media/File:Ceph_components.svg>.

7 Zero To Hero Guide For Ceph Cluster Planning. <https://ceph.com/geen-categorie/zero-to-hero-guide-for-ceph-cluster-planning/> 02.01.2014 Luettu: 30.01.2019.

8 Steven J. Vaughan-Nichols. Ceph: Block Storage for the 21st Century. <https://medium.com/linode-cube/ceph-block-storage-for-the-21st-century-b59f3b62d4ca> 11.07.2019. Luettu: 30.03.2019.

9 Data Placement Overview. <http://docs.ceph.com/docs/mimic/rados/operations/data-placement/> Luettu: 30-.03.2019.

10 Guillaume Chenuet. Ceph - Look around the CRUSH Map. <http://yauuu.me/ride-around-ceph-crush-map.html> 12.07.2016. Luettu: 07.05.2019.

11 Vmware. <https://www.techopedia.com/definition/16053/vmware> Luettu: 31.03-.2019.

12 Difference between vSphere, ESXi and vCenter. <http://www.mustbegeek.com/difference-between-vsphere-esxi-and-vcenter/> 24-.08.2012. Luettu: 31.03.2019.

13 Tom Collins. Virtual Servers vs. Physical Servers: Which Is Best? <https://www.atlantech.net/blog/virtual-servers-vs.-physical-servers-which-is-best> 10.03.2016. Luettu: 02.04.2019.

14 What is File Level Storage vs. Block Level Storage?<https://stonefly.com/resources/what-is-file-level-storage-vs-block-level-storage> Luettu: 02.04.2019.

15 David Marshall, Stephen S. Beaver, and Jason W. McCarty. VMWare ESX Performance Optimization. <http://www.ittoday.info/Articles/VMWare_ESX_Performance_Optimization.htm> 2008. Luettu: 07.05.2019

16 vSAN Concepts. <https://docs.vmware.com/en/VMware-vSphere/6.5/com.vmware.vsphere.virtualsan.doc/GUID-ACC10393-47F6-4C5A-85FC-88051C1806A0.html> 18.04.2018. Luettu: 07.05.2019

17 Enable IT Transformation with Managed Services. <https://www.netapp.com/us/services/managed-services.aspx> Luettu: 10.04-.2019.

18 Dedicated block storage for hybrid clouds and server clusters. <https://iweb.com/block-storage> Luettu: 10.04.2019.

19 Nitheesh Poojary. Understanding Object Storage and Block Storage Use Cases. <https://cloudacademy.com/blog/object-storage-block-storage/> 12.03.2019. Luettu: 10.04.2019

20 Logan G. Harbaugh. Block, file and object storage interfaces enable integration. <https://searchstorage.techtarget.com/feature/Block-file-and-object-storage-interfaces-enable-integration> 05.2018. Luettu: 10.04.2019.

21 Dell EMC Isilon Scale-Out NAS Storage <https://shop.dellemc.com/en-us/Solve-For/STORAGE-PRODUCTS/Dell-EMC-Isilon-Scale-Out-NAS-Storage/p/ISILON-Scale-Out-NAS-Storage> Luettu: 10.04.2019.

22 EMC Isilon Scale-Out Storage and Vmware vSphere. <https://www.emc.com/collateral/software/technical-documentation/h10555-sz-isilon-vmware-vsphere-sizing.pdf> 02.2012. Luettu 11.04.2019.

23 Kenneth Hui. OpenStack Compute For vSphere Admins, Part 1: Architectural Overview. <https://cloudarchitectmusings.com/2013/06/24/openstack-for-vmware-admins-nova-compute-with-vsphere-part-1/> 24.06.2013. Luettu: 11.04.2019.

24 Michael Gugino. Deploying Cinder with Multiple Ceph Cluster Backends. <https://medium.com/walmartlabs/deploying-cinder-with-multiple-ceph-cluster-backends-2cd90d64b10> 12.10.2016. Luettu: 11.04.2019.

25 Kuvalähde <https://raw.githubusercontent.com/ceph/ceph/master/doc/images/stack.png>.

26 Federico Lucifredi. Choosing The Right Storage For Your OpenStack Cloud. <https://people.redhat.com/%7Eflucifre/talks/Choosing%20the%20right%20storage%20for%20your%20OpenStack%20cloud%20webinar.pdf> 12.07.2017 Luettu: 18.04.2019.

27 Sébastien Han. Ceph RBD and iSCSI. <https://www.sebastien-han.fr/blog/2017/01/05/Ceph-RBD-and-iSCSI/> 05.01.2017. Luettu: 18.04.2019.

28 Kuvalähde. <https://www.heise.de/select/ix/2018/6/1527816560127139> 06.2018.

29 Ceph iSCSI Gateway. <http://docs.ceph.com/docs/mimic/rbd/iscsi-overview/> Luettu: 07.05.2019.

30 Ceph-sähköpostilistan poiminta käyttäjäkokemuksista NFS:n ja iSCSI-GW:n kanssa. <http://lists.ceph.com/pipermail/ceph-users-ceph.com/2015-July/003178.html> Luettu: 07.05.2019.

31 Sander van Vugt. Integrate Ceph object storage in a VMware vSphere environment. <https://searchvmware.techtarget.com/tip/Integrate-Ceph-object-storage-in-a-VMware-vSphere-environment> Luettu: 18.04.2019.

32 Understanding Ceph journals. <https://access.redhat.com/articles/1237243> 10.05.2016. Luettu: 30.04.2019.

33 Sam McLeod. Benchmarking IO with FIO. <https://web.archive.org/web/20170407115509/https://smcleod.net/benchmarking-io/> 29.04.2017. Luettu: 18.04.2019.

34 ‘dd’ command in Linux. <https://www.geeksforgeeks.org/dd-command-linux/> Luettu: 18.04.2019.

35 Random vs. Sequential explained. <https://blog.open-e.com/random-vs-sequential-explained/> Luettu: 19.04.2019.

36 FIO How-To. <http://git.kernel.dk/cgit/fio/plain/HOWTO> Luettu: 19.04.2019.

37 Sebastian Kirsch. Understanding congestion in vSAN. <https://www.justvirtualthings.com/understanding-congestion-in-vsan/> 27.06-.2018. Luettu: 30.04.2019.

38 Esimerkkiskripti. <https://github.com/cernceph/ceph-scripts/blob/master/tools/crush-reweight-by-utilization.py> Luettu: 30.04.2019.

39 Stephen McElroy. Recovering from a complete node failure. <https://ceph.com/planet/recovering-from-a-complete-node-failure/> 06.05.2017. Luettu: 30.04.2019.

40 FRA1 Block Storage Issue. <https://status.digitalocean.com/incidents/8sk3mbgp6jgl> 02.04.2018. Luettu: 30.04.2019.

41 New in Luminous: BlueStore. <https://ceph.com/community/new-luminous-bluestore/> 01.09.2017. Luettu: 30.04.2019.

42 Suomeen hankitaan uusi supertietokone – 37 miljoonan investointi. <https://www.tivi.fi/uutiset/suomeen-hankitaan-uusi-supertietokone-37-miljoonan-investointi/0ad6f2a4-0cbb-36c8-9024-d56a5c467111> 14.11.2018. Luettu: 30.04.2019.

43 Kimmo Koski. Competiveness through infrastructure and skillful people. <https://fdcf.fi/wp-content/uploads/datacenter-seminar-13_2_2019-Kimmo-Koski.pdf> 13.02.2019. Luettu: 30.04.2019.

44 Elmer <https://research.csc.fi/-/elmer?redirect=https%3A%2F%2Fresearch.csc.fi%2Fsoftware-details%3Fp_p_id%3D101_INSTANCE_qvEXMgE3ObVa%26p_p_lifecycle%3D0%26p_p_state%3Dnormal%26p_p_mode%3Dview%26p_p_col_id%3Dcolumn-2%26p_p_col_pos%3D1%26p_p_col_count%3D4%26p_r_p_564233524_categoryId%3D53926%26p_r_p_564233524_resetCur%3Dtrue> Luettu: 07.05.2019.

45 Optimal Integrated Ceph Solutions at Petabyte Scale. <http://go.qct.io/solutions/software-defined-storage/qxstor-red-hat-ceph-storage-edition/> Luettu: 07.05.2019.

46 Leo Kosola. Mitä sinun pitäisi tietää big datasta, datanlouhinnasta ja datafuusiosta? <https://yle.fi/aihe/artikkeli/2016/06/28/mita-sinun-pitaisi-tietaa-big-datasta-datanlouhinnasta-ja-datafuusiosta> 28.06.2016 Luettu: 07.05.2019.

47 Gerhard Sulzberger. Internal S3 Storage for Media Files on Top of Ceph Storage Technology. <https://www.runtastic.com/blog/en/internal-s3-storage-media-files-top-ceph-storage-technology/> 06.12.2017. Luettu: 07.05.2019.

48 Gartner Report: The Future of Object Storage. <https://www.netapp.com/us/forms/campaign/gartner-future-obj-storage.aspx> Luettu: 30.04.2019.