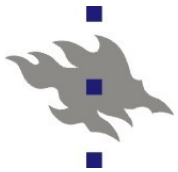
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Linux-ylläpito, kevät -10

Jani Jaakkola
jjaakkol@cs.helsinki.fi
http://www.cs.helsinki.fi/u/jjaakkol/lyp2010/

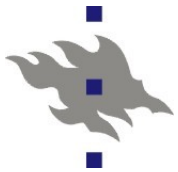
2. työasemaluentojen kalvosetti

- Lohkolaitteet
  - Loogiset ja Fyysiset
- */etc/fstab* – Tiedostojärjestelmien listaus
- RAID
  - RAID konfiguraatiovaihtoehdot
  - RautaRAID
- *tmpfs*
- Tiedostojärjestelmät
  - Ext3, Ext4,
  - XFS, JFS
  - FAT, ISO, UDF, HFS+
  - Brtfs
- Kernelin loogiset lohkolaitteet
  - Loopback, SoftaRAID, Devicemapper ja kryptaus
  - LVM
- Quota



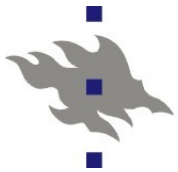
lohkolaitteet

- Laitetiedosto, josta pääse käsiksi levyn sisältöön
  - Fyysiset oikeat laitteet
    - Kovalevyt (*/dev/sda0)*
    - RAID
    - FC: fibre channel
    - optiset mediat (*/dev/scd0*)
  - Virtuaaliset laitteet
    - SoftaRAID: */dev/md0*
      - monta oikeaa laitetta muodostaa virtuaalisen RAID0, RAID1 tai RAID5 laitteen
    - LVM2
      - Monta oikeaa laitetta luo virtuaalisen laitteen, jonka koko voi vaihtua ja jossa fyysiset laitteet voivat vaihtua
    - Device mapper ja loopback
      - Laitteet, joiden alla on jotain älykkyyttä
      - Kryptaus, tavallinen tiedosto lohkolaitteena, multipath



/etc/fstab

- Listaa asennuksen kiinteät tiedostojärjestelmät
  - Irroitettavat mediat nykyään liitetään hal-daemonin kautta
  - Ainakin juuritiedostojärjestelmä
    - Usein */home* on erillinen tiedostojärjestelmä
    - Verkkotiedostojärjestelmät (NFS, SMB)
  - <laite> <liitoskohta> <tj:n tyyppi> <optiot> <milloin ajetaan fsck>
  - Laite voi olla suora polku laitetiedostoon, tj:n UUID tai LABEL
    - Verkkotiedostojärjestelmillä omat syntaksit
  - Tiedostojärjestelmän tyyppi (usein vain ”auto”)
  - Optiot tiedostojärjestelmäkohtaisia, tosin:
    - ro: readonly, nosuid: ei kunnioiteta suid-bittejä, nodev: ei tueta laitetiedostoja
    - user: annetaan käyttäjälle oikeus liittää tiedostojärjestelmä kutsumalla */bin/mount*:ia suoraan

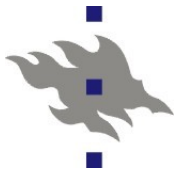


Työkaluja

- */sbin/blkid*
  - Listaa tunnettuja lohkolaiteiden tyypit ja attribuutit
    - LABEL, UUID
- */proc/partitions*
  - Kernelin näkemät lohkolaitteet, niiden koot ja laitenumerot
- */bin/mount, /bin/umount*
  - lohkolaitteiden liittäminen tiedostojärjestelmään
- */sbin/hdparm*
  - Työkalu IDE, SATA ja SCSI kovalevyjen hallintaan
- */sbin/smartctl, hddtemp*
  - Smart-protokollalla kovalevy osaa kertoa omasta tilastaan
- Partitiotaulueditorit
  - *fdisk, sfdisk,* jne
- */sbin/fsck*
  - Tiedostojärjestelmäriippumaton työkalu eheystarkastuksiin
  - Kutsuu varsinaista */sbin/fsck.<fstyyppi>* työkalua

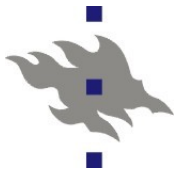
Swap ja tmpfs

- Keskusmuistia on rajallinen määrä
  - Muistin loppuessa muistisivuja voidaan kirjoittaa levylle
  - Harvoin käytetyt sivut voidaan siirtää levylle
- Swap-tila on partitio tai tiedosto (mahdollisesti useampia), joka on varattu tähän tarkoitukseen
- levyvälimuistin sivuja ei talleteta swappiin
  - Niille on tila varattu itse tiedostojärjestelmästä
- tmpfs on tilapäistiedostoille tarkoitettu fs
  - Tieto on normaalisti keskusmuistissa, mutta voidaan tallettaa swappiin mikäli tiedostoja ei käytetä tai keskusmuistia tarvitaan muuhun käyttöön
- Hibernoidessa järjestelmän tila talletaan swapille
- Swap-tilan loppuessa linux tyypillisesti jumiutuu tai kaatuu
- Työkalut */sbin/mkswap*, */sbin/swapon, /sbin/swapoff, /bin/free*



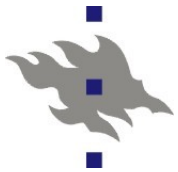
RAID

- Redundant Array of Inexpensive Disks
  - Kohta ”inexpensive” ei pidä paikkansa: palvelinten RAID-levyt ovat kalliita verrattuna kuluttujalevyihin
  - RAID-pakka: useista fyysisistä levyistä koottu looginen levy
- RAID0
  - Monta fyysistä levyä kytketty yhdeksi isoksi levyksi
  - Jos yksi levy hajoaa, koko RAID-pakka on rikki
- RAID1: mirroring
  - Kaksi saman kokoista levyä kytketty yhdeksi loogiseksi
  - Kaikki data kirjoitetaan molemmille levyille: toinen levy voi hajota, mutta data pysyy tallessa
- RAID5
  - 3 tai useampia levyjä kytketty isoksi loogiseksi levyksi
  - Yksi levyistä voi hajota, mutta data pysyy tallessa
- RAID6
  - RAID5, mutta ylimääräinen hotspare-levy



RautaRAID

- Erillinen ohjain: vaatii ohjainkohtaisen ajurin
- Tarvitaan myös erilliset ohjainkohtaiset hallintatyökalut RAID-pakkojen konfigurointiin
  - Tyypillisesti myös yksinkertainen BIOS-työkalu, jonka käyttö edellyttää koneen uudelleenkäynnistystä
- Tarvitsee BIOS-ajurin, jotta RAID-levyltä voi ladata käynnistyslataajan, kernelin ja initrd:n
- Valvontatyökalu(t)
  - Jos (ja kun) RAID-pakasta hajoaa levy, täytyy siitä toimittaa tieto, että uusi levy saadaan tilalle
- Patrol read
  - Ohjaimen firmware lukee itsekseen levyä läpi, etsien virheitä
- Työkalut firmware-päivityksiin
  - Valitettavasti näitä joutuu tekemään

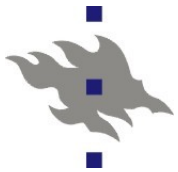


Tiedostojärjestelmän valinta

- Tiedostojärjestelmän valintaperusteet tärkeysjärjestyksessä:
  - 1. Luotettavuus
  - 2. Ominaisuudet
  - 3. Tehokkuus
- Miksi näin?
  - Tiedostojärjestelmän tehokkuudella ei ole juuri väliä
    - Järjestelmän tehokkuusongelmat ovat muualla
  - Tiedostojärjestelmän ominaisuuksillakaan ei ole niin väliä, jos bitit eivät pysy tallessa
    - Reiserfs oli hyvä ja tehokas, mutta jos rauta teki bittivirheitä, koko tj saattoi päätyä bittien taivaaseen
- Linuxin levyvälimuisti kirjoittaa:
  - Metadatan (hakemistot, tiedostojen attribuutit) levylle 5s kuluessa muutoksista
  - Tiedoston sisällön levylle 30s kuluessa muutoksista

TJ:n Ominaisuuksia

- Toipuminen kaatumisista
  - Koneen kaatuessa väistämättä bittejä hukkuu
    - Linuxin voi oletusasetuksilla pitää tietoa 30s levyvälimuistissa
  - Journaloiva tiedostojärjestelmä takaa, että tiedostojärjestelmä voidaan palauttaa konsistenttiin tilaan
  - *Fsck* ja tarkistussummat auttavat bittivirheiden havaitsemisessa ja korjaamisessa
- ACL (Access Control Lists)
  - Oikeuksien monipuolisempi hallitseminen
- Hakemistoindeksit isoille hakemistoille
  - */usr/lib*
- Tiedostojärjestelmän koon säätäminen
  - Online tai Offline tilassa
  - lohkolaitteen kokoa vaihtamalla tai lisäämällä partitioita
- Fragmentoitumisen välttäminen



Ext3

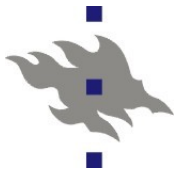
- Vanha ja stabiili tiedostojärjestelmä
  - Perinteinen unix tj: inode:t alustetaan tiedostojärjestelmää luodessa ja niitä on vakiomäärä
  - Kolme eri tapaa journaloida:
    - Journal: Metadata **ja** tiedostojen sisältö kierrätetään kirjoitettaessa journalin kautta
      - Tj on aina täysin konsistentti, mutta kaikki data joudutaan kirjoittamaan kahteen kertaan
    - Ordered: Vain metadata kierrätetään journalin kautta
      - Tiedoston sisältö synkronoidaan levylle kun inode viedään journaliin
      - Kompromissi ja oletusasetus
    - Writeback: vain metadata kierrätetään journalin kautta
      - Tiedostojen sisällön säilymisestä ei ole mitään takuita

Ext3 ominaisuuksia

- Vanha ja luotettava
  - Inode-rakenteen ansiosta kestää todennäköisemmin korruptiota
- Tuki hakemistoindekseille
- Ei erityisen nopea
  - Toisaalta ei tuhlaa CPU-aikaa
  - Erityisesti isojen tiedostojen poisto on hidasta
- Tiedostojärjestelmätarkastukset kestävät kauan
  - Laitoksen 2T kokoisilla tiedostojärjestelmillä lähes 2h
- Osaa kasvattaa itseään, myös online-tilassa
- Yhteensopiva vanhan ext2:n kanssa
  - Ext2 + journal
- Ei tue poistettujen tiedostojen palautusta
  - ext2:lla tämä oli vielä mahdollista

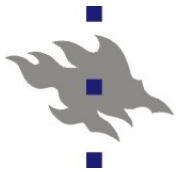
Ext3 työkalut

- */sbin/mke2fs*
  - Formatoi ext3-tiedostojärjestelmän
  - Hidas: inode-listojen alustukseen menee aikaa
- */sbin/tune2fs*
  - Tiedostojärjestelmän attribuuttien listaus ja säätö
- */sbin/e2fsck*
  - Tiedostojärjestelmän eheystarkastus
- */sbin/resize2fs*
  - Tiedostojärjestelmän kasvatus ja pienennys
- */sbin/dump* ja */sbin/restore*
  - Varmistuskopiointi suoraan tiedostojärjestelmän lohkolaitteelta (marginaalisesti tehokkaampaa)
  - Vaarallista: ei mitään takuita varmistuskopion ehjyydestä
- Joukko debuggaustyökaluja



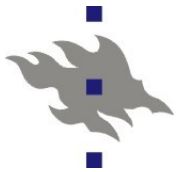
Ext4

- Ext3:n seuraava versio
  - Julistettu stabiiliksi kernelissä 2.6.28
- Extents:
  - Levytilan varaus ei pohjaudu enää lohkobittikarttoihin
  - Nopeuttaa levytilan varausta ja vapautusta isoilla tiedostoilla
  - Ei ole yhteensopiva ext3:n kanssa
- Viivästetty levytilan varaus
  - Varaus tehdään vasta kun data kirjoitetaan levylle
- Pysyvät levytilan etukäteisvaraukset
- Ajoaikainen fragmentoinnin poisto
- Tarkastussummat journalissa
- Aikaleimat nanosekunnin tarkkuudella
- Nopeammat tiedostojärjestelmätarkastukset
- Yli 32000 alihakemistoa yhdessä hakemistossa



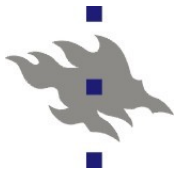
XFS

- SGI:n unixeille kehitetty journaloiva tiedostojärjestelmät
- Journaliin talletetaan tiedostojärjestelmäoperaatioita kokonaisten blokkien sijaan (kuten ext3)
  - Journal on pienempi ja tehokkaampi
  - Ei tue datan journalointia
- Tehokkaampi varausjärjestelmä
  - dynaamisesti varatut inodet
- Ajoaikainen fragmentoinnin poisto
- Ajoaikainen kasvattaminen ja kutistaminen
- Tuki reaaliaikaoperaatioille
  - API, jolla sovellus voi pyytää suoraa pääsyä levylle levyvälimuistin ohi taatulla siirtonopeudella



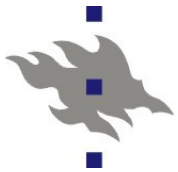
JFS

- IBM:n unixeille ja OS/2:lle kehitetty journaloiva tiedostojärjestelmä
  - Journaloi ainoastaan metadan kuten XFS
- Hakemistopuut
- Extents
- Jne..
- Ei kehitetä kovin aktiivisesti



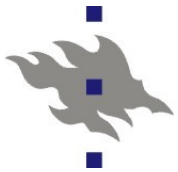
ISO9660

- iso9660-standarditiedostojärjestelmä optiselle medialle
  - Kirjoitetaan kerran, luetaan monta kertaa
  - *genisofs* -käyttäjätason ohjelmisto tiedostojärjestelmän luontiin
  - Erilliset ohjelmistot tj:n polttamiseen optiselle medialle
  - Nykyään käytössä toimivat GUI-ohjelmistot, jotka piilottavat ikävät yksityiskohdat (k3b)
- Variantit:
  - Rockridge: lisäattribuutit unix-bittien ja pitkien tiedostonimien tallettamiseen medialle
  - Joliet: MS:n laajennus pitkille tiedostonimille
- Nykyään cdrom:ien liittäminen tapahtuu GUI-ohjelmista hal:in avustuksella



FAT

- File Allocation Table
- MS:n tiedostojärjestelmä DOS-käyttöön
  - Teknologiaa suoraan 80-luvun 16-bittisistä laitteista
- Tuhlaa tilaa, erityisesti isoilla tiedostojärjestelmillä
- Ei tukea >2G kokoisille tiedostoille
- Fragmentoituu ja hajoaa helposti
- VFAT lisäsi tuen pitkille tiedostonimille
- Valitettavasti edelleen ainoa kaikkien käyttöjärjestelmien yhteisesti tukema tiedostojärjestelmä
  - Käytetään muistitikuilla ja korteilla

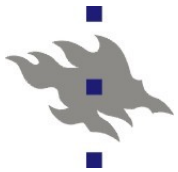


UDF

- Universal Disk Format
- Tiedostojärjestelmä optisille medioille
  - Myös ylikirjoittaville
- Korvaa ISO9660-standardin
- DVD-levyillä
  - Myös video DVD:t ja bluray
- Mahdollistaa suoran kirjoituksen optiselle RW-medialle
- Jonain kauniina päivänä ehkä korvaa FAT:in

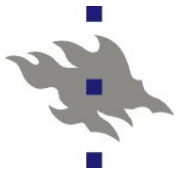
NTFS, HFS+

- NTFS on Windows-käyttöjärjestelmien natiivi tiedostojärjestelmä
  - Linuxille on kaksi NTFS-ajuria:
    - Vanhempi kernelin sisäinen ajuri, joka on toimiva read-only käytössä
    - FUSE-ajuri, jota voi käyttää myös NTFS tiedostojärjestelmälle kirjoittamiseen ilman pelkoa tj:n hajoamisesta
  - Ei tue Unix-tiedostobittejä eikä tunne tiedoston omistajan käsitettä
  - Linuxista käsin ei pääse muokkaamaan NTFS tiedosto-oikeuksia
- HFS+
  - Mac OS X:n natiivi tiedostojärjestelmä
  - Tuntee Unix-bitit ja tiedoston omistajan
  - Toimiva Linux-tuki, tosin ilman journalointia



Brtfs

- Kesken oleva kehitysprojekti
  - Takana Oracle
- Moderni tiedostojärjestelmä
- Tarkastussummat sekä datalle että metadatalle
- Ajoaikainen:
  - Lohkolaitteiden lisäys ja poisto
  - Defragmentointi
  - Blokkilaitteiden kuorman tasaus (tietoa siirtämällä)
  - Kompressointi
  - Tilannekuvat (snapshot)
- read-only median käyttäminen tj:n alustuksessa
- Tiedostojärjestelmän transaktiot
- Suunnitellut ominaisuudet
  - RAID5 tuki tiedostojärjestelmän sisällä
  - Inkrementaalivarmistukset
  - Automaattinen duplikaattiblokkien poisto

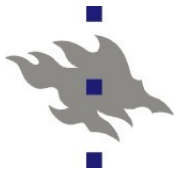


Loopback -lohkolaite

- Loopback -lohkolaitteen avulla tavallisen tiedoston saa näyttämään lohkolaitteelta
- Hyödyllinen erityisesti CDROM ja DVD-levyimageja käytettäessä
- Korppuja ei nykyään käytetä enää mihinkään, mutta korppuimagelta usein käynnistetään laitteita, esim. BIOS-päivityksiä varten

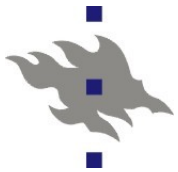
Linux softaRAID

- Kernelin sisäinen RAID-toteutus
- Tukee RAID0, RAID1 ja RAID5 tasoja
- initrd:n ansiosta myös juuritiedostojärjestelmä voi olla RAID-levyllä
- BIOS **ei** osaa käynnistää linuxia softaRAID-levyltä
- Linux osaa tallettaa luotujen RAID-pakkojen metadatan RAID-partitioille
  - Pakka saadaan automaattisesti käyttöön käynnistyessä
  - Levyjen fyysisellä sijainnilla ei ole väliä
- Hallintatyökaluna */sbin/mdadm*
  - */sbin/mdadm --create*: pakan luonti
  - */sbin/mdadm --detail:* pakan tiedot
  - */sbin/mdadm --add*: lisää levyn pakkaan
  - */sbin/mdadm --remove*: poistaa levyn pakasta
  - */proc/mdstat*: Kernelin mielipide RAID-pakan tilasta



Rauta vs. SoftaRAID

- Aidot RAID-ohjaimet ovat kalliita
- Käytännössä SoftaRAID on hidas
  - Teoriassa näin ei pitäisi olla
- Hajoava RAID-ohjain vie mukanaan koko pakan
  - Toisaalta myös levyohjain voi hajota
  - Multipath-viritykset mahdollistavat saman RAID-pakan käyttämisen useamman ohjaimen kautta
- SoftaRAID ei toivu kaatumisista siististi
  - Linux ei tiedä, ehdittiinkö viimeisetkin bitit kirjoittaa kaikille levyille
  - RautaRAID on akuilla suojattua
- SoftaRAID voi erehtyä luulemaan levyä rikkinäiseksi
  - jokin ihan muu ongelma aiheuttaa timeoutin
- Linuxin SoftaRAID ei aktiivisesti lue levyjä
  - Ei patrol read -tukea



Devicemapper ja kryptaus

- Kernel-moduli, kirjasto ja työkalut virtuaalisten lohkolaitteiden luontiin ja hallintaan
  - Virtuaalilohkolaitteet koostuvat yhdestä tai useammasta fyysisestä laitteesta
- Työkalu */sbin/dmsetup*
- Kryptattujen tiedostojärjestelmien alustus ja käyttö
  - */sbin/cryptsetup*
  - Tukee LUKS salausavainten hallintastandardia
    - Useampia salausavaimia samalle partitiolle
    - Esim. Järjestelmäylläpitäjälle oma salasana ja käyttäjälle oma
  - Ilmeisesti LUKS-levyille pääsee käsiksi myös windowsista

LVM: Logical Volume Management

- LVM mahdollistaa levyjen lisäämisen, vaihtamisen ja poistamisen levyjärjestelmästä, ilman tiedostojärjestelmien uudelleenalustusta
  - Ei tue vikasietoisuutta: se täytyy toteuttaa RAID:in avulla
- LVM niputtaa yhden tai useamman fyysisen levyn yhdeksi levynideryhmäksi (Volume Group)
- Levynideryhmälle luodaan levyniteitä (Volume)
  - Levyniteet käyttäytyvät kuten partitiot: levyniteelle tehdään tiedostojärjestelmä, jota lopulta käytetään
- Levynideryhmät tukevat erilaisia operaatioita:
  - Levyniteiden luonti ja poisto
  - Levyniteiden kasvatus ja kutistus
    - Vaatii tukea tiedostojärjestelmältä
  - Fyysisten levyjen poisto ja lisäys
  - Levynideotokset (snapshot) tietyn täsmällisen tilan varmistuskopiointia tai talletusta varten

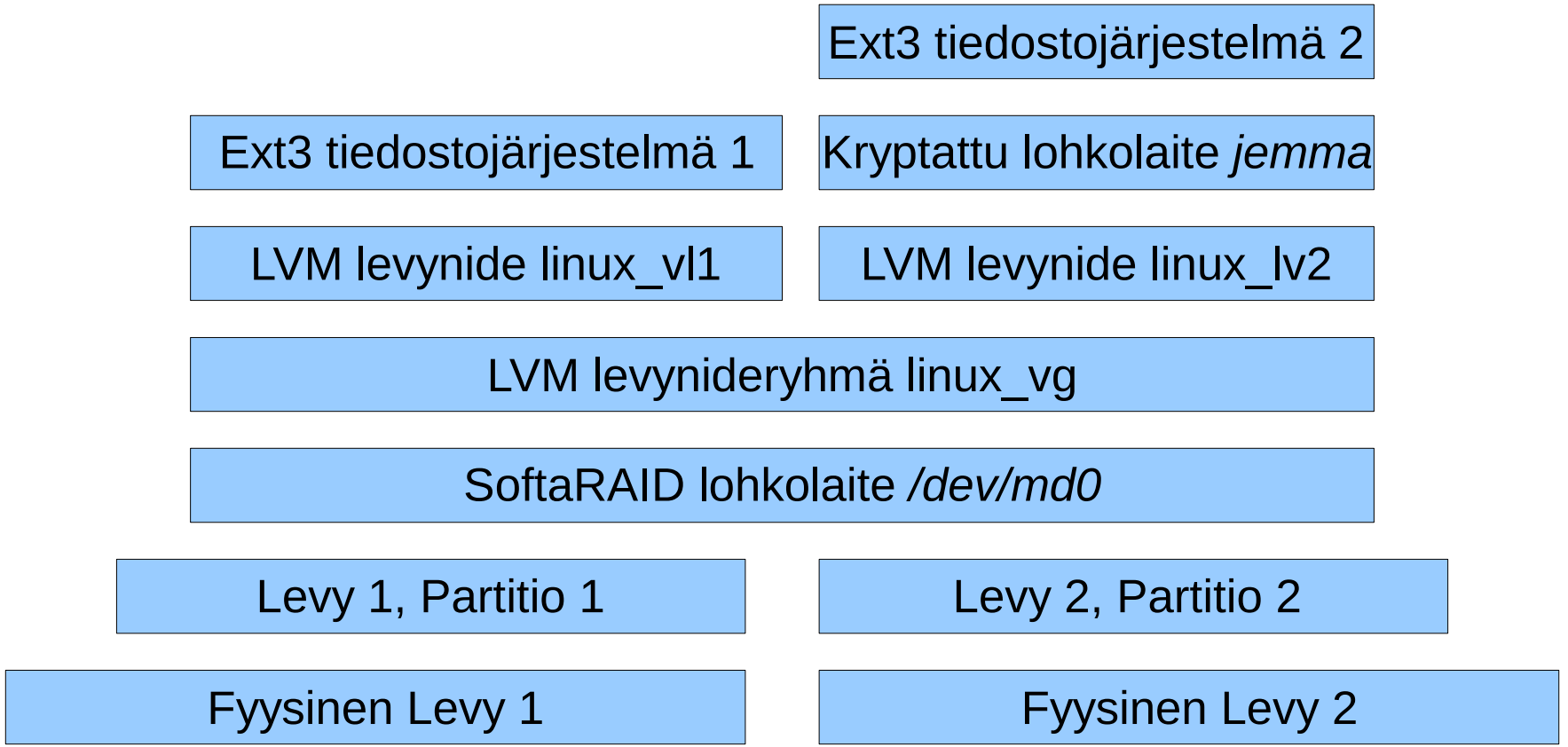
LVM työkalut

- */usr/sbin/pvcreate*
  - Fyysisen levyn alustus LVM-käyttöön
- */usr/sbin/vgcreate, /usr/sbin/vgremove*
  - Levynideryhmän luonti ja poisto
- */usr/sbin/vgextend, /usr/sbin/vgreduce*
  - Fyysisten levyjen lisäys ja poisto levynideryhmään
- */usr/sbin/lvreate, /usr/sbin/lvremove*
  - Levyniteen luonti ja poisto
  - Otoksien (snapshot) luonti ja poisto
- */usr/sbin/lvextend, /usr/sbin/lvreduce*
  - Levyniteen koon kasvatus ja pienennys
- */usr/sbin/lvs, /usr/sbin/lvdisplay*
  - Leviniteiden tila, levyniteen attribuutit

LVM: sudenkuoppia

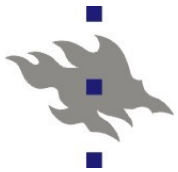
- Käynnistyslataajat eivät osaa lukea levyniteitä
  - */boot* täytyy ainakin sijoittaa oikealle fyysiselle partitiolle
- Initrd:lle täytyy asentaa LVM-työkalut, jotta kernel osaisi löytää LVM levyniteelle sijoitetun juuritiedostojärjestelmän
  - Tämän pitäisi nykyä tapahtua distribuutioiden puolesta automaattisesti
- Levyjen lisäys ja poisto ja vikasietoisuus pitäisi oikeasti olla tiedostojärjestelmän ongelma
  - Nyt ylläpitäjä joutuu manuaalisesti rakentamaan RAID:in, levyniteet ja komentomaan tiedostojärjestelmää
  - Brtfs voi olla tulevaisuuden ratkaisu

Esimerkkikonfiguraatio



Quota: käyttäjien levynkäyttö

- Oletusarvoisesti Linux-tiedostojärjestelmät eivät laske käyttäjien levynkäyttöä
  - */etc/fstab* tiedostoon pitää lisätä optio *quota*
- Linuxissa on erillinen järjestelmä levynkäytön laskentaan
  - Quotatieto talletetaan erilliseen tiedostoon tiedostojärjestelmän juureen (*/aquota.user* ja */aquota.group)*
  - Quota-järjestelmä päivittää näitä tiedostoja sitä mukaan käyttäjät luovat tiedostoja ja käyttävät levytilaa
  - Quota-tiedostoille kannattaa lisätä ext3-tiedostojärjestelmän bitti +j, jotta kaikki tiedostoihin tehtävät muutokset kiertäisivät journalin kautta
  - Quota-tiedostot täytyy erikseen alustaan
  - XFS-tiedostojärjestelmässä on sisäinen quota-toteutus
- Quotatyypit
  - soft-quotan voi ylittää tilapäisesti, hard-quotaa ei voi ylittää



Quota komentoriviltä

- */usr/bin/quota*
  - käyttäjän quotan listaus
- */usr/sbin/setquota*
  - käyttörajojen säätö
- */usr/sbin/repquota*
  - kaikkien käyttäjien quotat
- */usr/sbin/quotacheck*
  - quota-tiedostojen alustus
  - Laskee käyttäjän omistamat tiedostot ja niiden viemän tilan
  - Ei toimi luotettavasti, jos tiedostojärjestelmä on käytössä
  - Hidas isoilla tiedostojärjestelmillä
- */usr/sbin/quotaon* ja */usr/sbin/quotaoff*
  - Quota käyttöön ja pois käytöstä
- NFS tukee quota-kyselyjä myös verkon yli
  - Tästä lisää myöhemmin verkkoluennoilla