M-grid: Linux Suomen ensimmäisessä tuotannollisessa Grid-ympäristössä

Arto Teräs <arto.teras@csc.fi>

Linux & Open Source -koulutus

Hotel Kämp, Helsinki, 1. marraskuuta 2005

Sisältö

- **CSC ja materiaalitutkimuksen grid (M-grid)**
- **Valmis paketti vai oma ratkaisu?**
- **Asennus ja päivitykset**
- **Ylläpito yhteistyönä — voiko se toimia?**
- **Käyttäjäkokemuksia**
- **Grid-yhteiskäyttö**
- **Tietoturvahaasteet**
- **Yhteenveto**

Tieteen tietotekniikan keskus CSC

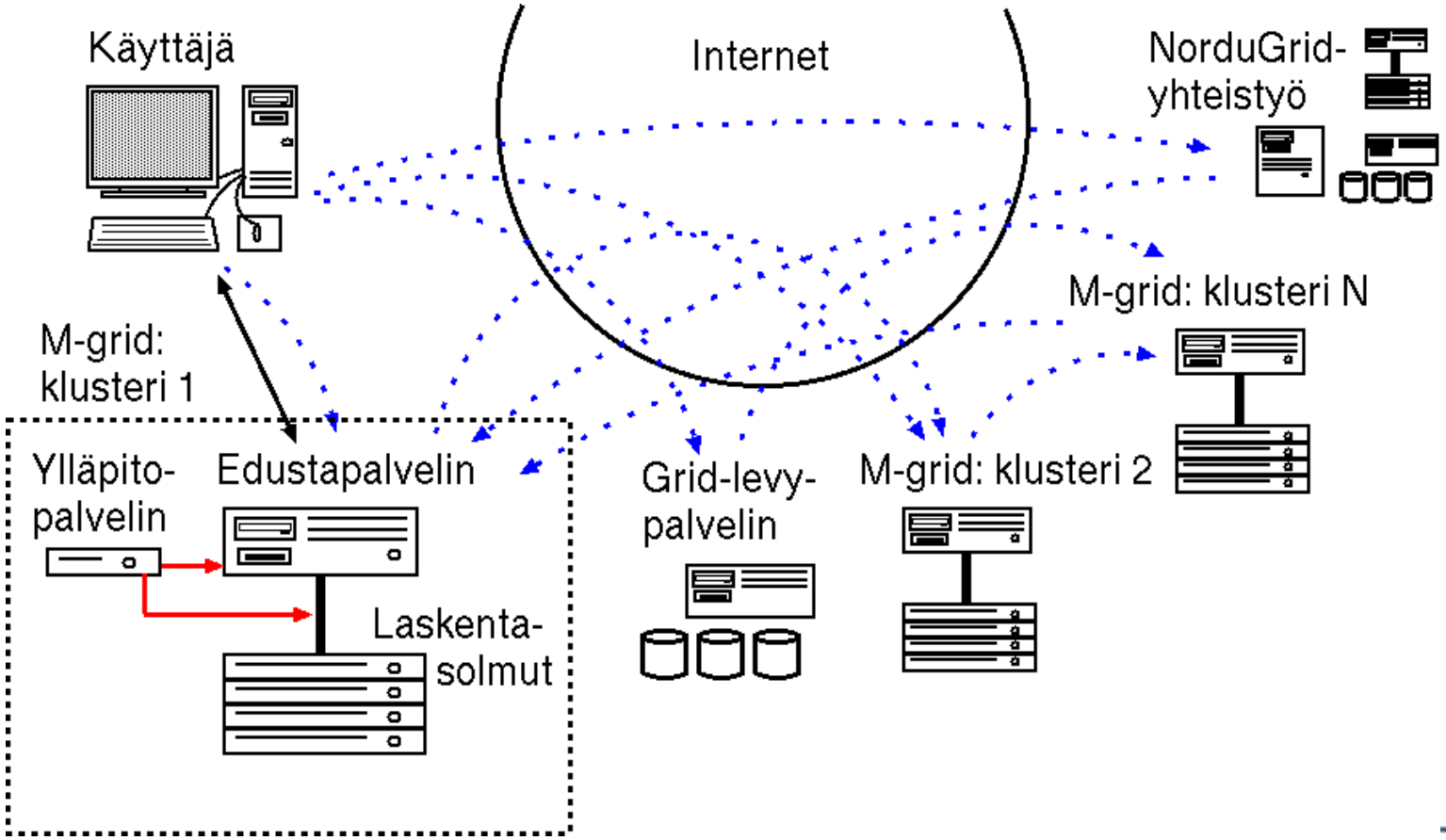
- **Tehtävä: tutkimuksen ja opetuksen kansallisen tason IT- palvelut sekä infrastruktuurin kehittäminen ja ylläpito**
- **Palvelualueet:**
  - Funet-palvelut
  - Laskentapalvelut
  - Sovellusten käyttöpalvelut
  - Tietohallinnon järjestelmäpalvelut
  - Tieteen tietotekniikan asiantuntijapalvelut
- **Asiakkaat: korkeakoulut ja tutkimus-laitokset ja niiden tietotekniikkaa hyödyntävä henkilökunta**
- **Omistaja: opetusministeriö**

Materiaalitutkimuksen grid (M-grid)

- **Tavoite: Laskentakapasiteettia lähinnä fysiikan ja kemian tutkijoiden tarpeisiin**
- **Seitsemän yliopiston, Fysiikan tutkimuslaitoksen ja tieteen tietotekniikan keskus CSC:n yhteishanke**
  - Kumppaneina lähinnä osastot, ei yliopistojen atk-keskukset
- **Rahoitus Suomen akatemialta ja osallistuvilta yliopistoryhmiltä**
  - Rahoitushakemus marraskuussa 2003, käyttöönotto lokakuussa 2004
- **Ensimmäinen suuri hanke Suomessa, jossa grid-teknologiaa otetaan tuotantokäyttöön**
- **Alusta: Linux-pohjainen PC-klusteriympäristö**

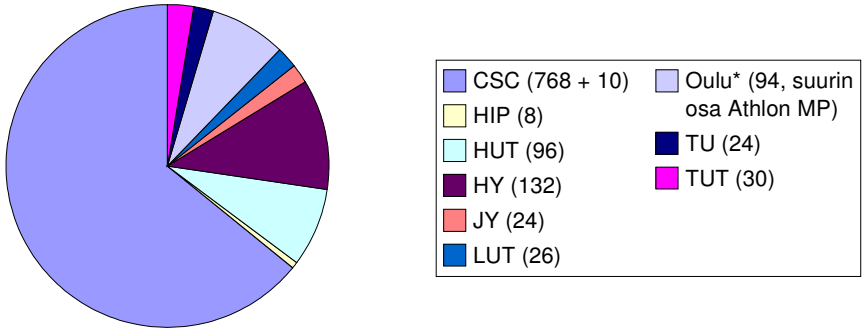
CSC

Grid-ympäristö



Laitteisto

- **Kymmenen keskenään eri kokoista klusterilaitteistoa**
  - Kahden AMD Opteron -suorittimen laskentasolmut (HP DL145): 1.8-2.2 GHz, 2-8 GB muistia, 80-320 GB paikallista levyä
  - Edustapalvelin (HP DL585): 1-2 TB jaettua levytilaa
  - Verkko 2 x Gbit Ethernet + etähallintaverkko + ylläpitopalvelin
- **Laskentasolmuissa yhteensä 778 (CSC) + 434 (yliopistot) suoritinta, teoreettinen kokonaislaskentateho 5 TFlop/s.**



Käyttöjärjestelmä ja grid-väliohjelmisto

- **NPACI Rocks Cluster Distribution**
  - Klustereihin tarkoitettu Linux-jakelu, pääkehittäjä San Diego Supercomputing Center
  - Perustuu Red Hat Enterprise Linuxin lähdekoodiin, mutta ei Red Hatin tuote
  - http://www.rocksclusters.org

- **SUN Grid Engine -eräajojärjestelmä**
  - Kunkin klusterin sisäinen laskentatöiden hallinta

- **NorduGrid ARC grid-väliohjelmisto**
  - Mahdollistaa laitteistojen yhteiskäytön siten, että väli-ohjelmisto valitsee verkosta vapaana olevan resurssin
  - http://www.nordugrid.org

Valmis paketti vai oma ratkaisu?

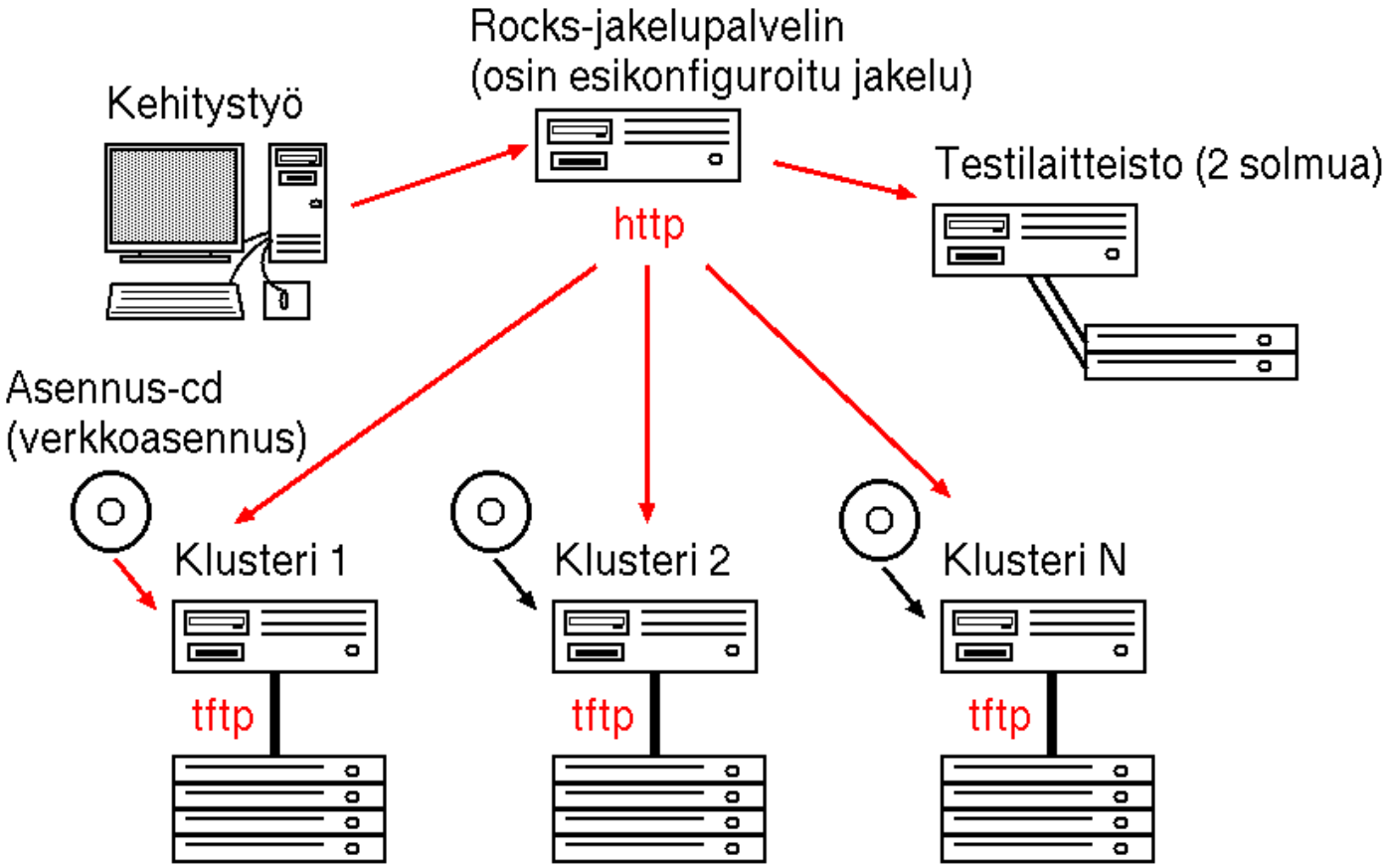
- **Linux oli helppo valinta — se on laskentaklustereissa jo johtava käyttöjärjestelmä**
- **Klusterihallintaan sekä kaupallisia että ei-kaupallisia vaihtoehtoja**
  - Valittiin Rocks, joka ei ole kaupallisesti tuettu mutta jolla on palkattu kehittäjätiimi ja laajahko käyttäjäkunta
  - Laitevalmistajien paketoimat ratkaisut parhaiten integroituja, toisessa vaakakupissa muokattavuus ja riippumattomuus
- **Kokonaan valmista pakettia ei olisi saatu kuitenkaan**
  - Open source -tuotteeseen voitiin perehtyä ja sovittaa omat sovellukset etukäteen riippumatta laitetoimittajan valinnasta
- **Luotettavuusvaatimus: varmatoiminen peruskäyttö, kokeellisempi grid-ympäristö**

CSC

Ylläpito M-gridissä

- **Tehtävät jaettu CSC:n ja paikallisten ylläpitäjien välillä**
- **CSC:n ylläpito:**
  - Käyttöjärjestelmän, eräajojärjestelmän, grid-väliohjelmiston ja tiettyjen varusohjelmakirjastojen ylläpito
  - Erillinen kahden laskentasolmun kokoonpano testausta varten
- **Paikalliset ylläpitäjät**
  - Paikallisen tutkijaryhmän sovellusten ylläpito, järjestelmän tilan seuranta, käyttäjätuki
- **Säännölliset tapaamiset ylläpitäjien kesken noin 2 kk välein, yhteinen postituslista**

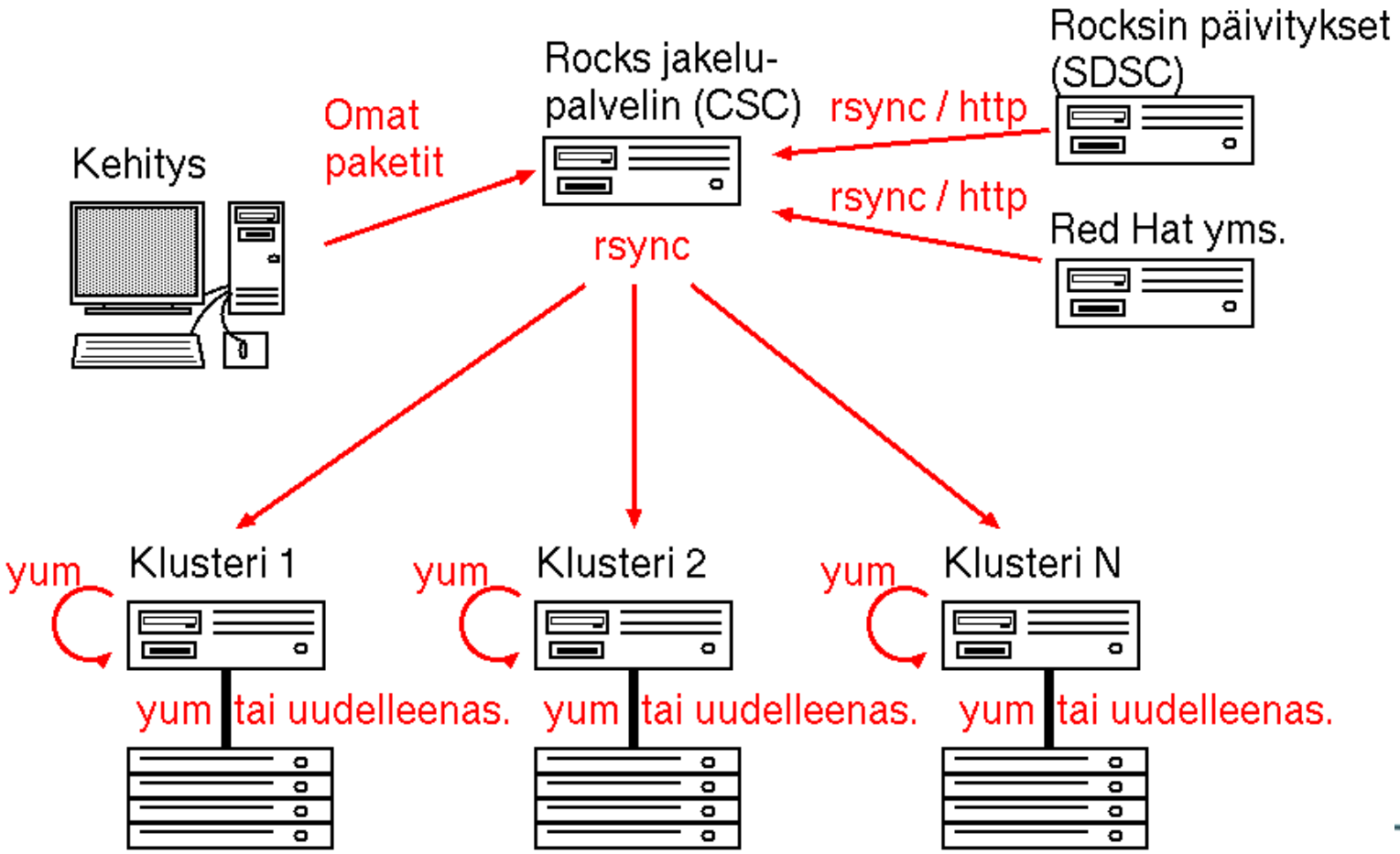
Asennuksen toteutus



Kokemukset käyttöönotosta

- **Laitetoimittajan tekninen henkilöstö suoritti laitteistoasennuksen**
- **CSC valmisteli käyttöjärjestelmän jakelupalvelimelle ja asennus-cd:n, paikalliset ylläpitäjät asensivat kukin oman klusterinsa**
- **Jakelun valmistelu vei oletettua enemmän aikaa**
  - Omiin räätälöinteihin saatiin tukea sähköpostilistan kautta suoraan Rocksin kehittäjiltä open source -yhteisölle tyypilliseen tapaan
- **Käyttöjärjestelmien asennus sujui hyvin**
  - Suurin osa saatiin valmiiksi alle päivässä, kahdessa suurimmassa klusterissa kului kaksi päivää
  - Yhdessä klustereista jouduttiin selvittämään outoa asennusongelmaa
- **Joitakin asetuksia erityisesti rinnakkaisajokirjastojen (MPI) osalta jouduttiin tekemään jälkeenpäin**

Päivitysten asentaminen



Rocksin vahvuudet ja heikkoudet

Hyvää:

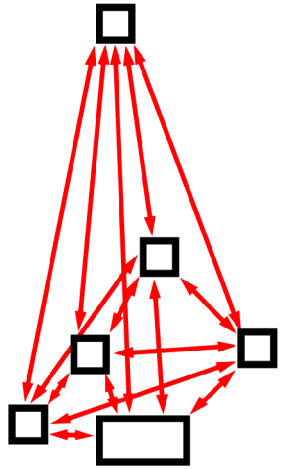
- **Helppo aloittaa, suunniteltu nimenomaan klustereihin**
- **Kätevät monitorointityökalut, moni asia toimii "out of the box"**
- **Suurimman osa toimittajista laitteistot RHEL-sertifioitu**
  => Rocks pitäisi myöskin toimia

Huonoa:

- **Rocks-tiimi ei julkaise omia tietoturvapäivityksiä eikä heiltä saa maksullista tukea**
  - Red Hatin source rpm -muodossa julkaisemat turvapäivitykset käyvät
- **Jakelun asennusta muokattaessa virheiden analysointi ja korjaus hankalaa**

Yhteistyönä toteutetun ylläpidon tavoitteet

- **Vahva keskitetty perusta paikallinen muokattavuus säilyttäen**
  - Uusi malli — perinteisesti Suomessa akateeminen suurteholaskenta on keskitetty CSC:lle
- **Helpompaa yliopistoille kuin rakentaa kokonaan oma klusteri**
  - Vaatii joka tapauksessa merkittävän työpanoksen niin CSC:ltä kuin paikallisilta ylläpitäjiltäkin
- **Hyödynnetään paikallisten ylläpitäjien erityisosaaminen**
  - Paikalliset ylläpitäjät tuntevat oman ryhmänsä sovellukset => parempi ja nopeampi käyttäjätuki



36 yhteistyöparia!

CSC

Ylläpito: hyvät kokemukset

- **CSC:n tuki on koettu hyödylliseksi**
  - Toisaalta paikallinen kontrolli (pääkäyttäjän oikeudet) mahdollistaa nopeatkin korjaukset ja on tärkeä psykologinen tekijä
- **Paikalliset ylläpitäjät ovat ottaneet hoitaakseen yhteisiä tehtäviä jotka osaavat hyvin — CSC ei ole tehnyt kaikkea**
- **Ryhmien välinen yhteistyö on tiivistynyt myös tutkimuksen osalta**
- **Laitteistot ovat lähellä käyttäjää**
  - Helppo kysyä neuvoa paikalliselta ylläpidolta, vähentää CSC:lle tulevia tukipyyntöjä
- **Suurin osa paikallisista ylläpitäjistä on myös itse käyttäjiä => suora palautekanava käytettävyydestä myös CSC:lle**

Ylläpito: huonot kokemukset

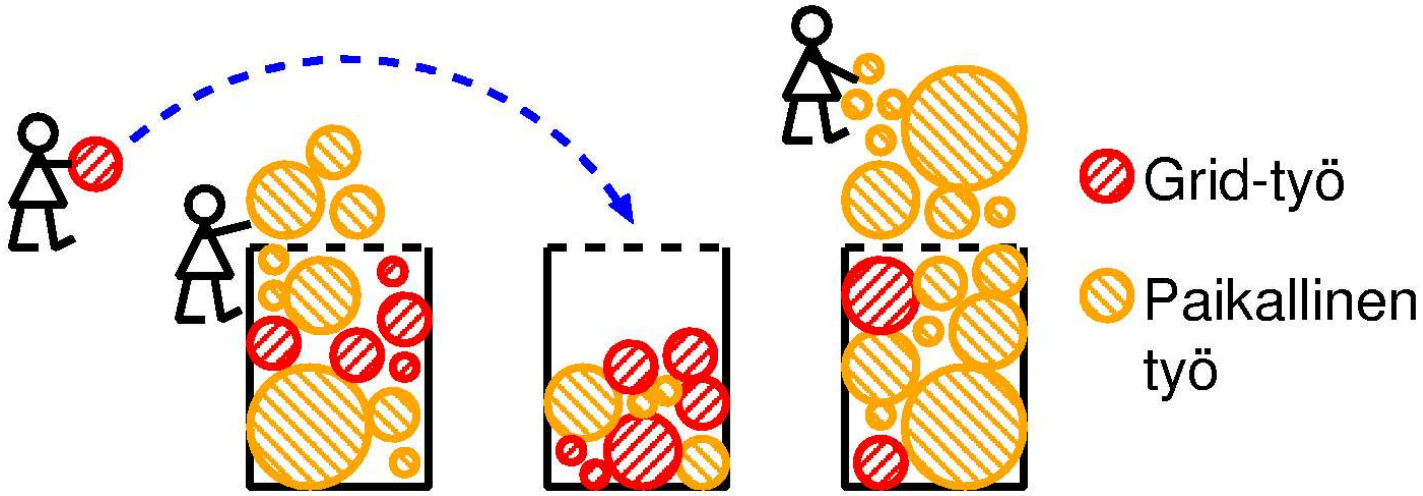
- **Sun Grid Engine v. 5.3 -eräajojärjestelmän konfigurointi**
  - Nykyinen versio 6.0 soveltuu klustereihin paremmin
- **Usein kysyttyjen kysymysten kerääminen Wikiin ei ole onnistunut vaan vastaukset jäävät postituslistalle**
  - Toisaalla Wiki-malli taas on ollut menestys, vrt. Wikipedia
- **Käyttäjädokumentaation puutteet**
  - Lähinnä kyseessä perinteinen henkilöresurssipula
  - Ohjeita voidaan kirjoittaa hajautetusti mutta niiden kerääminen vaatii keskitettyä koordinointia
- **Osa käyttäjistä ei saanut tarpeeksi tukea**
  - Kokemukset paikkakuntien kesken vaihtelevia: osa käyttäjistä ollut hyvinkin tyytyväisiä

Ensimmäisen vuoden käyttökokemukset

- **Käyttäjät löysivät koneet hyvin: muutaman kuukauden jälkeen keskimääräinen kuormitus yli 50%, nykyisin lähes 100%**
  - Linux oli tälle käyttäjäryhmälle ennestään tuttu ympäristö
- **Laitteistojen suorituskykyyn on oltu tyytyväisiä**
- **Luotettavuus ollut pääosin hyvä, vaikka ongelmilta ei ole täysin vältytty**
  - Edustakoneissa aluksi vakausongelmia, MPI-rinnakkaisajot herkkiä ympäristön muutoksille
- **Fortran-kääntäjän valinta aiheutti päänvaivaa**
  - GNU:n Fortran-kääntäjä toimii mutta tuottaa hidasta koodia
  - Osa sovelluksista yhteensopivia vain tietyn kääntäjän kanssa

Grid-yhteiskäyttö ja resurssien jako

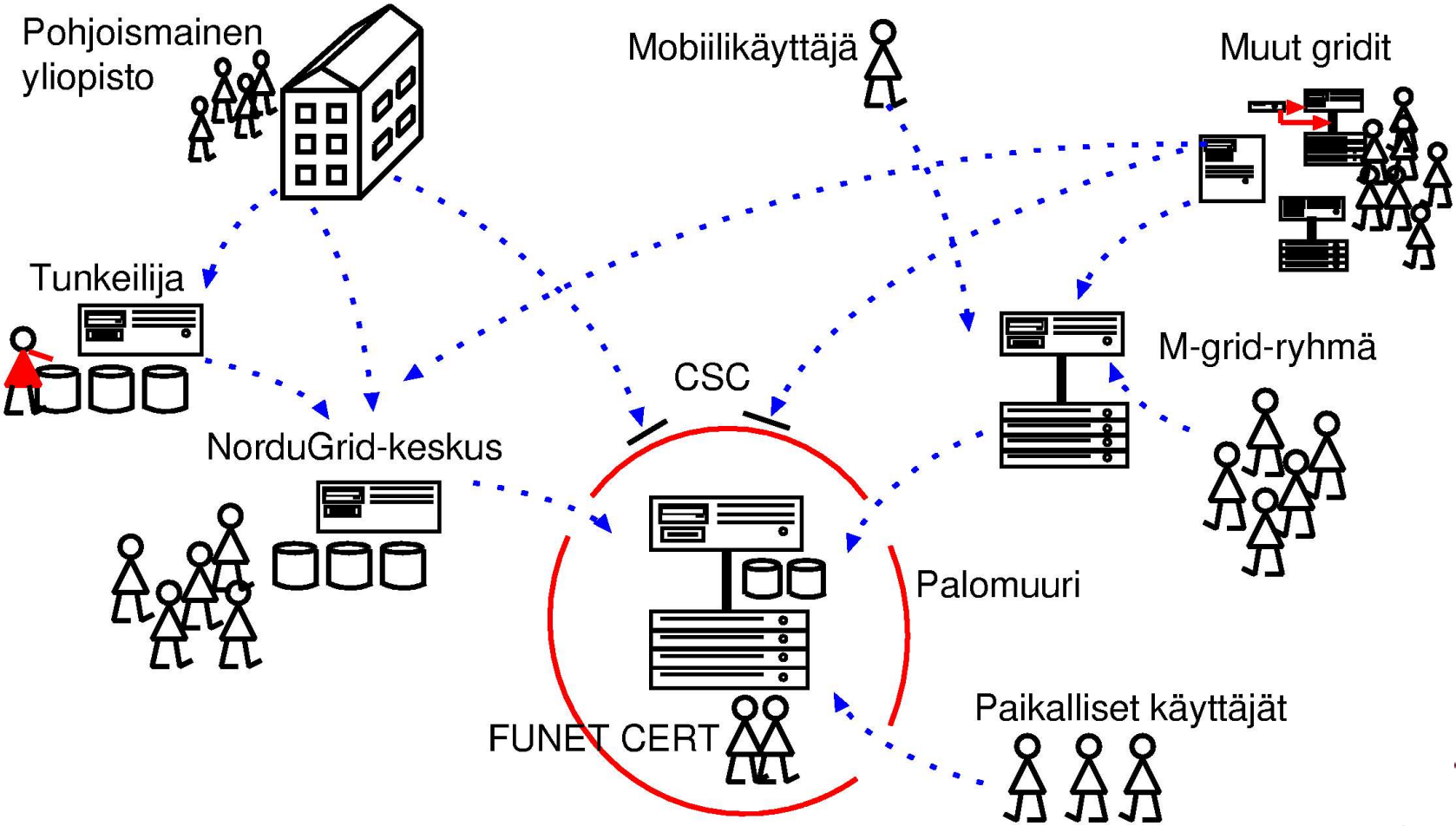
- **Käyttöpolitiikka: Laskentatöitä voidaan lähettää sekä paikallisesti että grid-liittymän kautta**
  - Resurssijako: paikalliset työt 80%, grid-työt 20%
- **Tavoitteena tyhjäkäynnin minimointi: tyhjiä solmuja ei pidetä varattuina (resurssijako dynaaminen)**



Grid-käyttökokemukset

- **Grid-käyttö alkoi elokuussa 2005**
  - Asennus viivästyi muiden kiireiden ja teknisten ongelmien takia
  - Ympäristö ei edelleenkään kaikilta osin valmis
- **Grid-ympäristön pitää olla parempi kuin paikallinen, muuten sitä ei käytetä!**
  - Pitkä jono omassa klusterissa ja tyhjä resurssi gridissä saattaa olla riittävä porkkana
- **Toistaiseksi vasta muutama grid-käyttäjä, aika näyttää kuinka hyvin grid-ympäristö otetaan vastaan**
- **Yhteistyömalli on toiminut: grid-projektit ovat aina muutakin kuin pelkkää tekniikkaa**

Grid-yhteistyö ja tietoturva



Tietoturvahaasteet grid-ympäristössä

- **Grid ylittää organisaatiorajat**

  => Keskinäinen luottamus avainasemassa!

- **Muutamia uusia uhkia ja kaikki vanhat laajalla vaikutuspiirillä**
  - Yksittäinen murrettu käyttäjätunnus edelleen helpoin tapa tunkeutua järjestelmään
  - Käyttäjätunnus gridissä antaa pääsyn suureen joukkoon koneita
- **Satojen käyttäjien järjestelmät ovat aina tietoturvariski**
  - Tietoturvaloukkauksia ei voi pitkällä tähtäimellä kokonaan estää: keskityttävä havaitsemaan murrot nopeasti
  - Selkeät toimintamallit tietomurtotapauksissa tarpeen

Tietoturvahaasteet (jatkoa)

- **Kaikki osapuolet eri tasoilla saatava mukaan**
  - Laskentakeskukset, yliopistojen atk-keskukset, paikalliset ylläpitäjät, CERT-ryhmät ja myös käyttäjät
  - Kansainvälinen yhteistyö
- **Vastuualueiden määrittely tärkeää luottamuksen muodostamiseksi**
  - Riskianalyysi
  - Käyttöpolitiikka ja käyttäjätunnusten hallinnointi
  - Toiminta tietomurtotapauksissa
  - Tietojen luottamuksellisuus ja yksityisyys

Yhteenveto

- **Ylläpito yhteistyönä voi toimia**
  - Henkilökohtaiset kontaktit tärkeitä — yhteiset tapaamiset paras tapa välttää sodat postituslistalla
- **Käyttäjätuki hajautetussa järjestelmässä parhaimmillaan erinomaista mutta vaatii erityishuomiota**
- **Valmista kokonaispakettia ei tarjolla: valittiin perusta jota voi laajentaa ja jonka päälle voi rakentaa**
- **Grid-projektit tiivistävät ryhmien välistä yhteistyötä riippumatta käytetystä teknologiasta**
- **Grid ylittää organisaatiorajat: ei mahdollista ilman keskinäistä luottamusta**

Lisätietoja

- **M-gridin kotisivu: http://www.csc.fi/proj/mgrid/**
- **Rocksin kotisivu: http://www.rocksclusters.org**
- **NorduGridin kotisivu: http://www.nordugrid.org**
- **Yhteyshenkilöt:**
  - Arto Teräs <arto.teras@csc.fi>
  - Kai Nordlund <kai.nordlund@helsinki.fi>
  - Olli-Pekka Lehto <oplehto@csc.fi> (Rocks)
  - Urpo Kaila <urpo.kaila@csc.fi> (tietoturva)
- **Kiitos! Kysyttävää?**