# M-grid: Linux Suomen ensimmäisessä tuotannollisessa Grid-ympäristössä

Arto Teräs <arto.teras@csc.fi>

HP Linux Forum

Messukeskus, Helsinki, 3.5.2005

# Sisältö

- **Materiaalitutkimuksen grid (M-grid): yleiskatsaus**
- **Kokemuksia käyttöönotosta**
- **NPACI Rocks Linux-jakelu monen klusterin ympäristössä**
- **Ylläpito yhteistyönä — voiko se toimia?**
- **Käyttäjäkokemuksia**
- **Tietoturvahaasteet**
- **Yhteenveto**

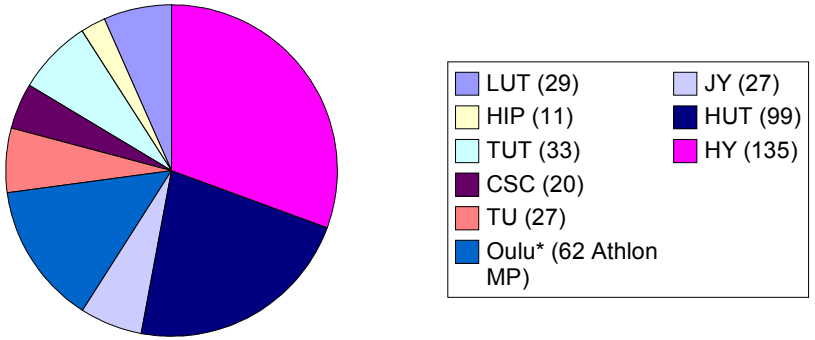
# Materiaalitutkimuksen grid (M-grid)

- **Tavoite: Edullista 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, käyttönotto lokakuussa 2004
- **Ensimmäinen suuri hanke Suomessa, jossa grid-teknologiaa otetaan tuotantokäyttöön**
- **Alusta: Linux-pohjainen PC-klusteriympäristö**


CSC

# Laitteisto

- **Yhdeksän 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 410 suoritinta, teoreettinen laskentateho 1.5 Tflops**

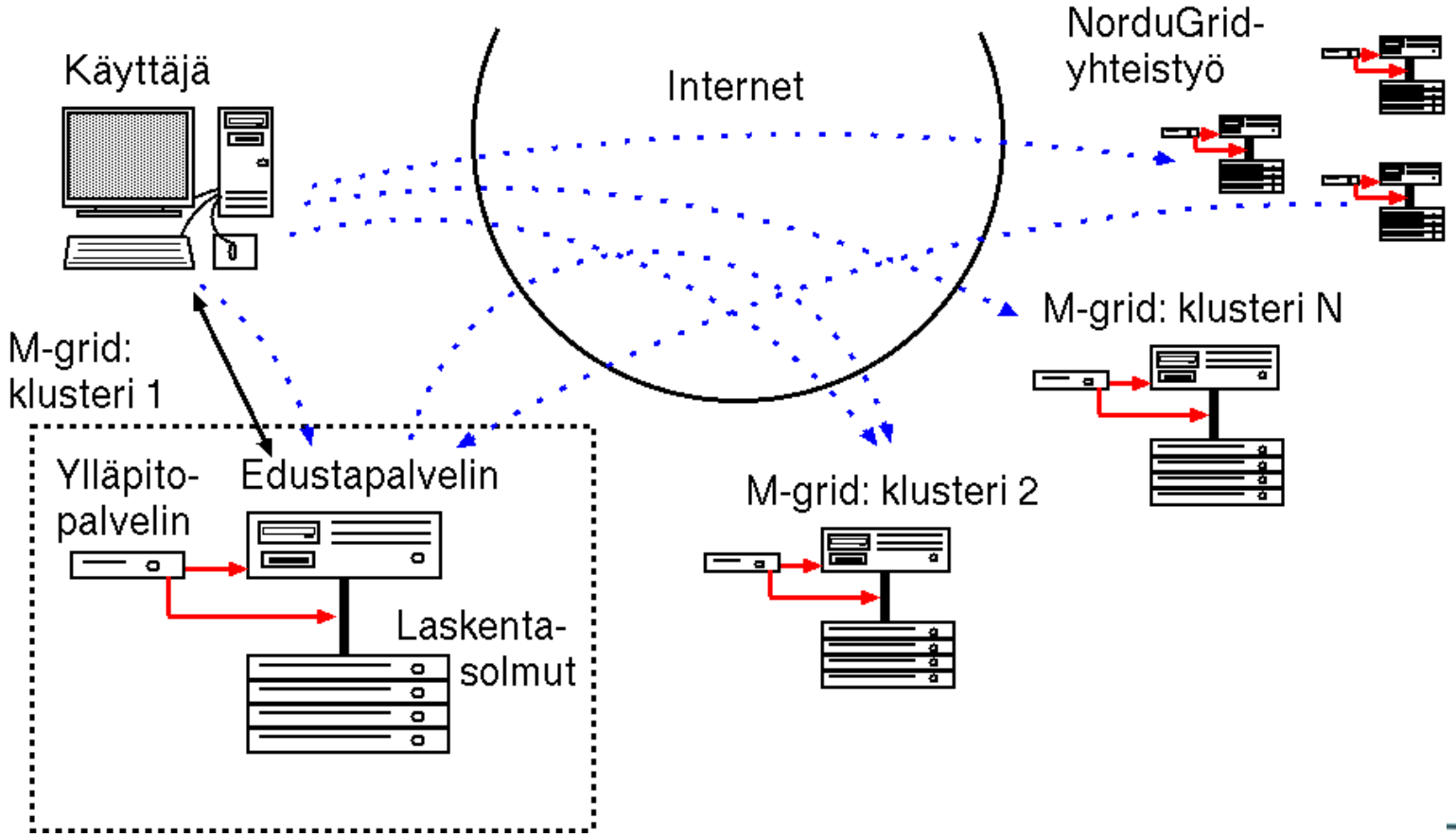


# Käyttöjärjestelmä ja grid-väliohjelmisto


- **NPACI Rocks Cluster Distribution**
  - Klustereihin tarkoitettu Linux-jakelu, pääkehittäjä San Diego Supercomputing Center
  - Perustuu Red Hat Enterprise Linux 3.0:n 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 väliohjelmisto valitsee verkosta vapaana olevan resurssin
  - http://www.nordugrid.org

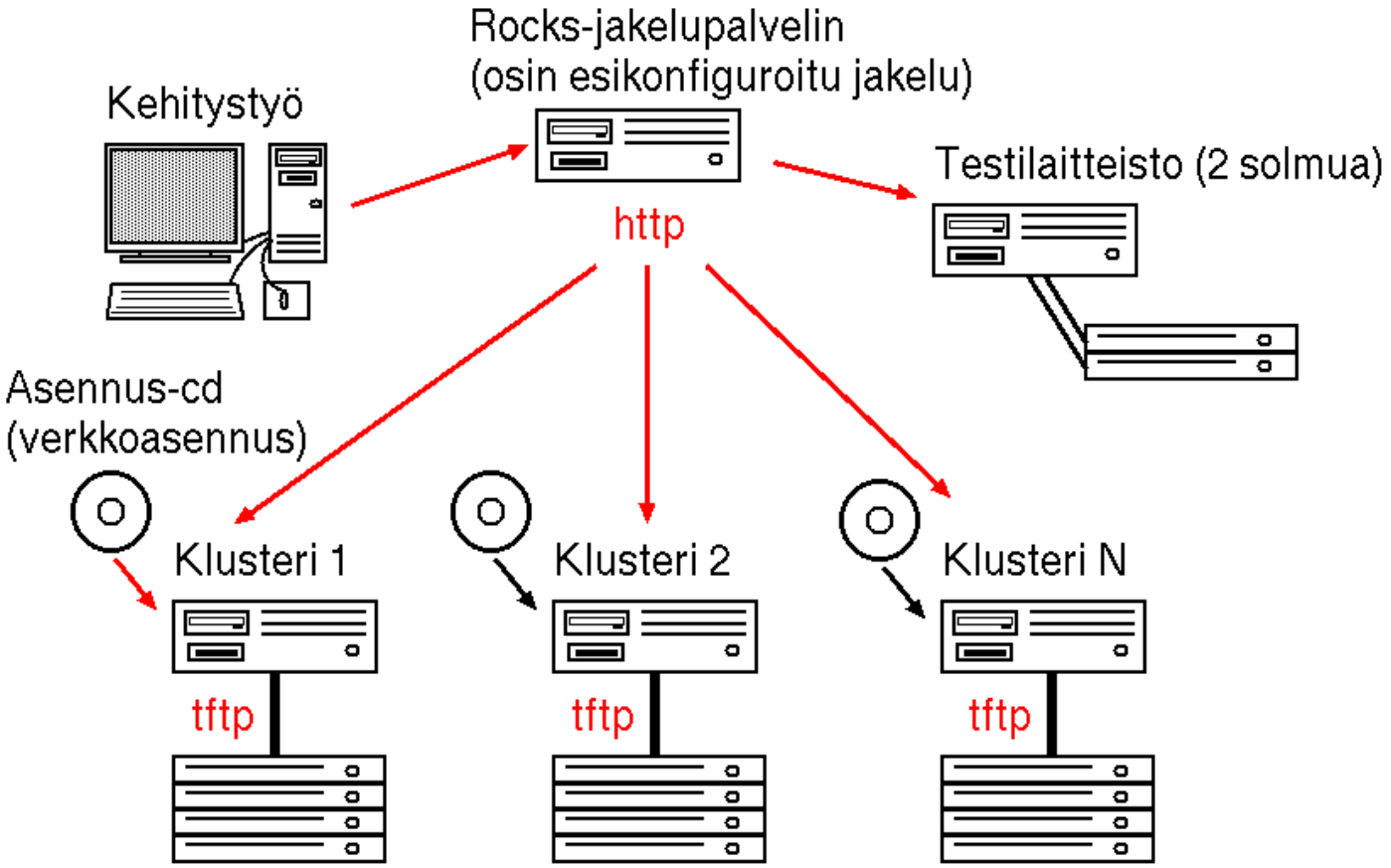

# Grid-ympäristö



# 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 n. 2 kk välein, yhteinen postituslista**

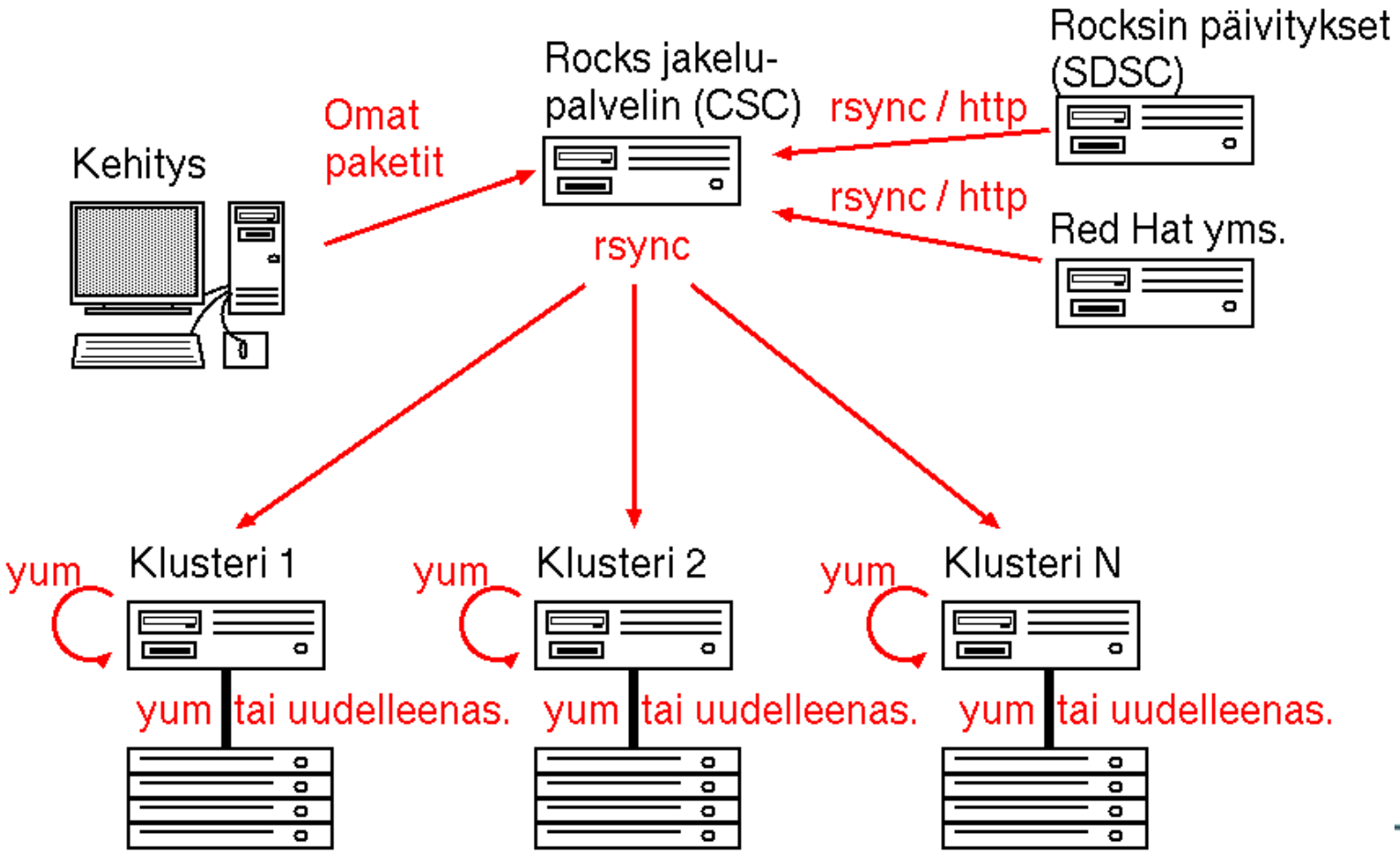
# Asennuksen toteutus



# Kokemukset käyttöönotosta

- **HP:n tekninen henkilöstö suoritti laitteistoasennuksen**
- **CSC valmisteli käyttöjärjestelmän jakelupalvelimelle ja asennus-cd:n, paikalliset ylläpitäjät asensivat kukin oman klusterinsa**
- **Jakelun valmistelu vei oletettua enemmän aikaa**
- **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:**

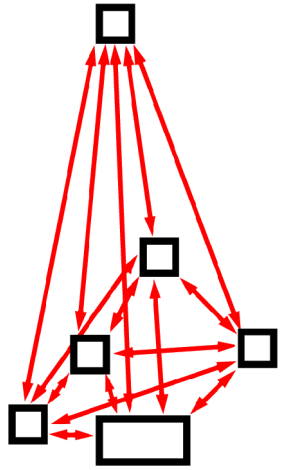
- **Helppo aloittaa, suunniteltu nimenomaan klustereihin**
- **Kätevät monitorointityökalut, moni asia toimii "out of the box"**
- **Suurimman osa toimittajista laitteistot RHEL-sertifioitu**
  **=> Rocks pitäisi myöskin toimia**

**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!**

# Ylläpito: hyvät kokemukset

- **CSC:n tuki on koettu hyödylliseksi**
  - Toisaalta paikallinen kontrolli (root-pääsyoikeus) 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**
  - Suurkonetausta, versio 6.0 huomioi myös klusterit paremmin
- **Usein kysyttyjen kysymysten kerääminen Wikiin ei ole onnistunut vaan vastaukset jäävät postituslistalle**
- **Käyttäjädokumentaation puutteet**
  - Lähinnä kyseessä perinteinen henkilöresurssipula
  - Dokumentaatiota voidaan kirjoittaa hajautetusti mutta sen kerääminen vaatii keskitettyä koordinointia
- **Osa käyttäjistä ei saanut tarpeeksi tukea**
  - Kokemukset paikkakuntien kesken vaihtelevia: osa käyttäjistä ollut hyvinkin tyytyväisiä

# Ensikuukausien käyttäjäkokemukset

- **Käyttäjät löysivät koneet varsin hyvin: muutaman kuukauden jälkeen keskimääräinen kuormitus yli 50%**
- **Laitteistojen suorituskykyyn on oltu tyytyväisiä**
- **Laskentasolmut toimineet erittäin luotettavasti, edustakoneissa ollut joitakin outoja ongelmia**
- **Ohjelmisto-ongelmat keskittyneet lähinnä Fortran-kääntäjään ja MPI-rinnakkaisajoihin**
  - MPI-töitä voi ajaa, mutta kaatuneiden prosessien varaamien resurssien automaattinen siivous ei toimi kunnolla
  - Useita eri Fortran-kääntäjiä saatavilla (GNU, Portland Group, Pathscale, Intel, ...), vaikea löytää yhtä joka tyydyttäisi kaikkia käyttäjiä

# Grid-käyttö

- **Valittu käyttöpolitiikka: Laskentatöitä voidaan lähettää sekä paikallisesti että grid-liittymän kautta**
  - Grid-töillä on paikallisia töitä suurempi prioriteetti 20%:ssa joka klusterista, lisäksi ne saavat täyttää kaikki vapaat solmut
- **Tilanne: Muussakin käyttönotossa riitti tekemistä joten grid-väliohjelmistojen käyttöönotto viivästyi ja sitä ollaan vasta nyt tekemässä**
  - 64-bittisen ympäristön ja Sun Grid Engine -eräajojärjestelmän sovittaminen yhteen ARCin kanssa aiheuttivat alkuhankaluuksia
  - Aika näyttää miten käyttäjät ottavat grid-ympäristön omakseen: tiivis yhteistyöverkosto toivottavasti auttaa
- **Yhteistyömalli on toiminut: grid-projektit ovat aina muutakin kuin pelkkää tekniikkaa**

# 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

# Yhteenveto

- **Ylläpito yhteistyönä voi toimia**
  - Henkilökohtaiset kontaktit tärkeitä — yhteiset tapaamiset paras tapa välttää sodat postituslistalla
- **Käyttäjätuki hajautetussa järjestelmässä voi olla erinomaista mutta myös huonoa: vaatii erityishuomiota**
- **Rocks soveltuu Linux-jakeluksi usean klusterin hajautettuun järjestelmään (muitakin vaihtoehtoja toki on)**
- **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?**