MTA SZTAKI

Department
of Distributed
Systems

A KOPI Plágiumkereső dinamikus skálázási kísérletei

Micsik András

Bevezetés

MTA SZTAKI
Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutatóintézet

- DSD = Department of Distributed Systems
  vagyis: Elosztott Rendszerek Osztály
  - szotar.sztaki.hu
  - kopi.sztaki.hu
  - lod.sztaki.hu
  - radio.sztaki.hu
  - szavazas.sztaki.hu

kopi.sztaki.hu

Terjedőben a plagizálás

Hungary President Schmitt quits in plagiarism scandal

Hungary's President Pal Schmitt says he is resigning, after being stripped of his doctorate over plagiarism.

Mr Schmitt, elected in 2010, said "my personal issue divides my beloved nation rather than unites it".

"It is my duty to end my service and resign my mandate as president," he told parliament.

Last week, Budapest's Semmelweis University revoked his 1992 award after finding that much of his thesis had been copied.

Mr Schmitt, 69, won gold medals for fencing at the 1968 and 1972 Olympic Games.

Mr Schmitt was an Olympic fencing champion before his rise in politics

Romanian prime minister accused of plagiarism

Allegations prompt questions about government's ability to tackle misconduct in academia.

Quirin Schiermeier

18 June 2012 | Updated: 20 June 2012

Romania's new government, still reeling from a misconduct scandal that forced its research minister to resign last month, has been hit by fresh allegations of plagiarism that strike at the very top.

Prime Minister Victor Ponta has been accused of copying large sections of his 2003 PhD thesis in law from previous publications, without proper reference. If the charges are substantiated, they could spark public pressure for Ponta to resign, say political insiders. The allegations are also raising fresh doubts about the government's ability to tackle corruption in the higher-education system.


Romanian Prime Minister Victor Ponta has been accused of copying large amounts of his PhD thesis from other sources without proper attribution.
XINHUA/PHOTOSHOT

German Defence Minister Guttenberg resigns over thesis


The German army was being restructured under Mr zu Guttenberg, so his departure leaves a gap in government

Time, CNN suspend Zakaria for plagiarism

Published August 10, 2012 /

CNN and Time magazine have suspended columnist and TV host Fareed Zakaria after he admitted to plagiarizing part of a New Yorker article on gun control.

Time issued a statement Friday that it had suspended Zakaria's column for a month while it investigates the matter further. CNN followed up later in the day announcing its own suspension, adding that it had removed a column by Zakaria on the same issue from its website.

A KOPI Plágiumkereső

- A kopi.sztaki.hu portál 2004-ben indult egynyelvű plágium keresési szolgáltatással (magyar és angol nyelveken)
- 2009: 10.000 regisztrált felhasználó
- 2011-ben a világon elsőként bemutattuk a fordítási plágiumkeresőt
  - Ez képes detektálni, ha valaki például az angol Wikipedia-ból lefordított bekezdéseket használ fel
  - Az új algoritmus számítási igénye nagyságrendekkel nagyobb, mint az egynyelvű plágiumkeresésé
- 2013: 25.000 regisztrált felhasználó

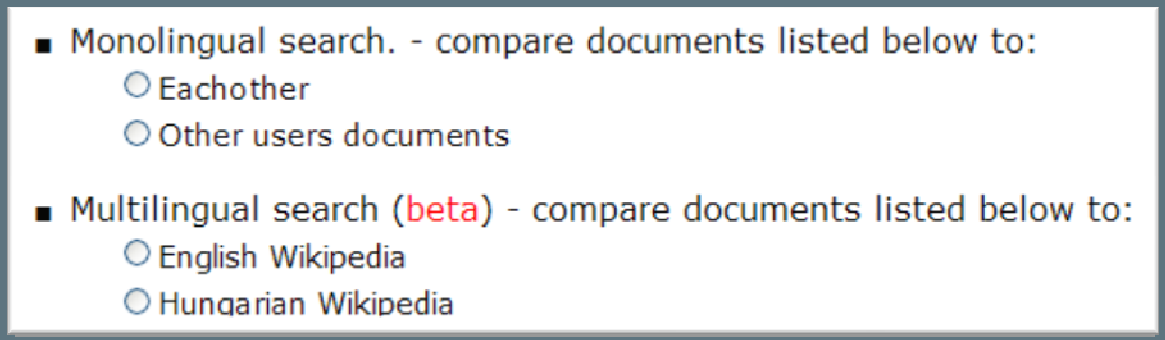
A KOPI Portál

Használat

1.

2.

3.



Eredmények

From: KOPI
Date: 2012.01.24.
Subject: 1 dokumentum összehasonlítása az angol Wikipédiával.

[Üzenet törlése]

2 hasonló mondatot talált a rendszer 3 Wikipédia cikkben:

1. Rövidítés (3)
   Rövidítésnek (latinul abbreviatura) nevezzük közszavak és tulajdonnevek rövidített formáit, melyek szinte kizárólag írott formában élnek, azaz amelyeket kiejtve teljes alakjukban használunk.

   - rövidítés és mozaikszó egy szó, kifejezés vagy név rövidített formája
     megjegyzés 1: rövidítésnek n
     formáit, melyek szinte kizáról
     teljes alakjukban használunk.

   (utca), km (kilométer), É (észak

   - (utca), km (kilométer), É (ész

1. Pete Seeger (7)
   Seeger was born in French Hospital, Midtown Manhattan, the youngest of three sons.

   - Pete Seeger Manhattan közepén, a Midtown-nak is hívott városrész francia kórházában született.

   His father, Charles Louis Seeger Jr. was a prominent musicologist, composer, and music professor.

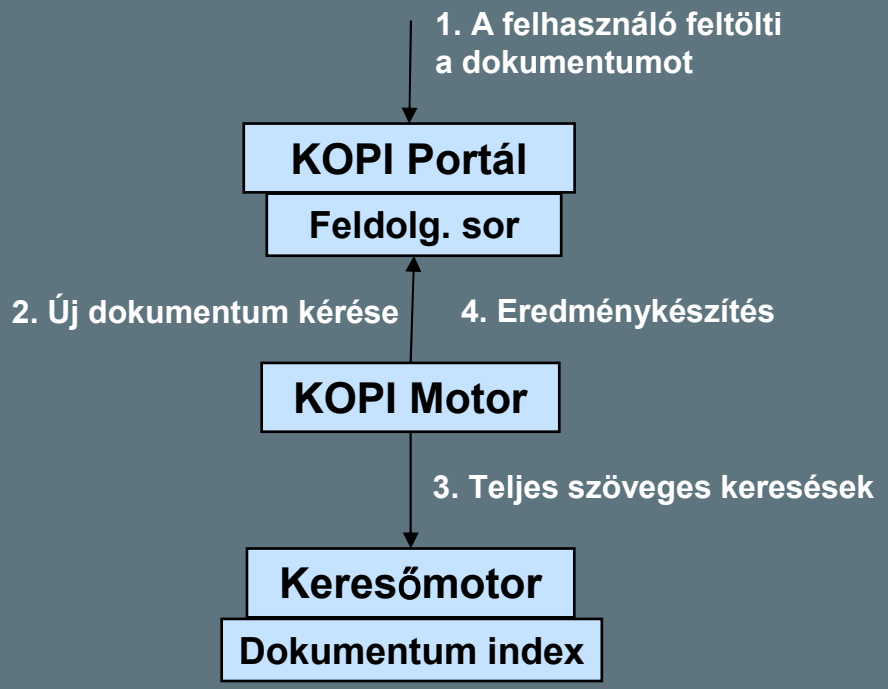
   - Apja, ifj. Charles Louis Seeger, zeneszerző és zenetudós volt, aki az elsők között vizsgálta mind az amerikai népzenét, mind a nem-európai gyökerekből fakadó zenét.

   His stepmother, Ruth Crawford Seeger, was one of the most significant female composers of the twentieth century.

   - Nevelőanyja, Ruth Crawford Seeger egyike volt a huszadik század legkiemelkedőbb női zeneszerzőinek.

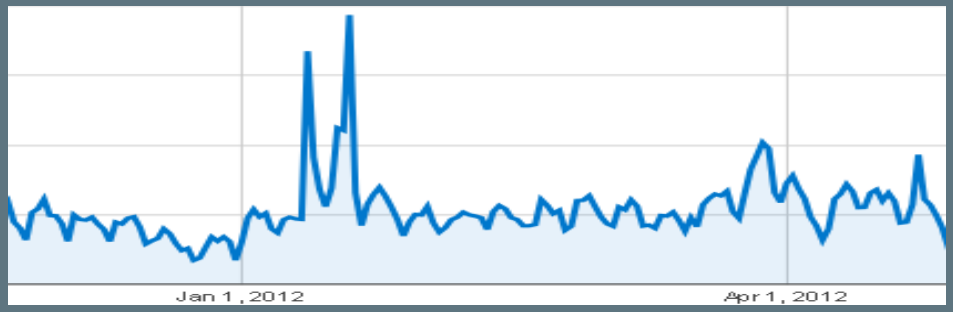
A plágiumkeresés folyamata

- A plágiumkeresés folyamata
  - A felhasználó feltölti a dokumentumot
  - A KOPI Motor új feladatot kér a Portáltól
  - A KOPI Motor feldolgozza a dokumentumot
  - Eközben teljes szöveges kéréseket ad ki a Keresőmotornak
  - A KOPI Motor összeállítja az eredményt és visszaküldi a Portálnak
  - A felhasználó értesítést kap, hogy az eredmény elkészült.
  - Az eredmény egy listát tartalmaz az esetlegesen másolt részekről és a plágium valószínűségéről

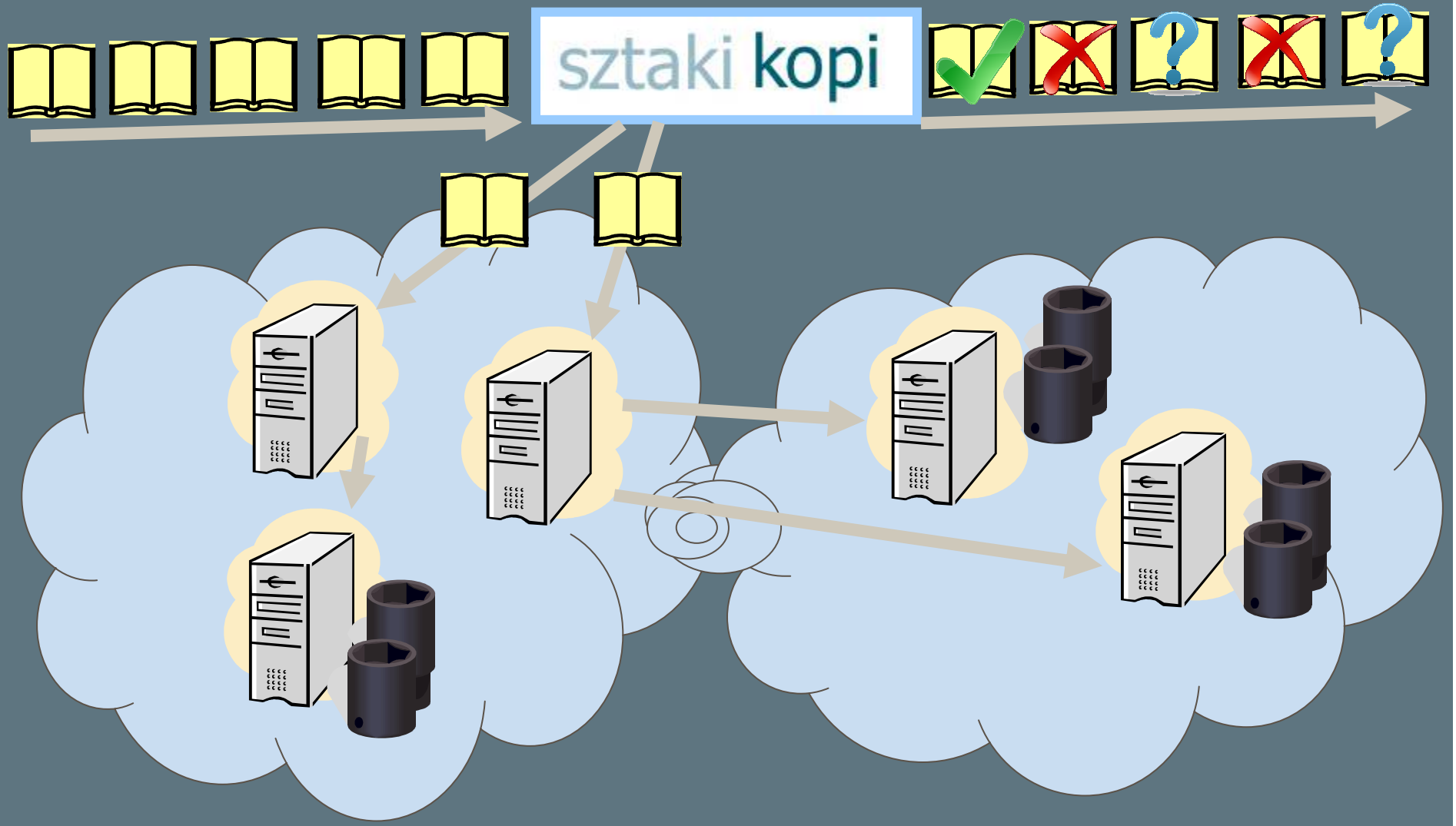


A kísérletről

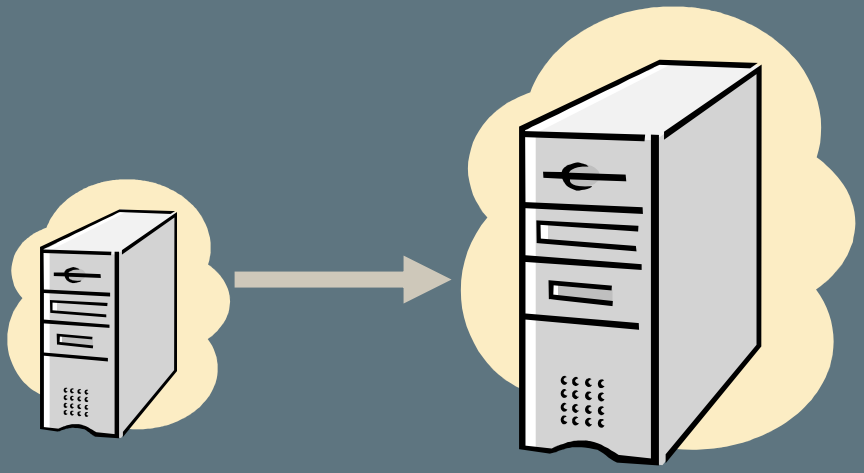
- A cél: stabil szolgáltatásminőség fenntartása
  - Egy dokumentum ellenőrzése igen változó, nagyban függ a mérettől, tipikusan 10-50 perc
  - Amikor túl sok dokumentum érkezik be, ez akár több napra is nőhet
- A kísérlet során
  - Modellezzük a tipikus felhasználói tevékenységet
  - Különféle skálázási módszereket mérünk heterogén felhő-szövetségekben



Skálázás

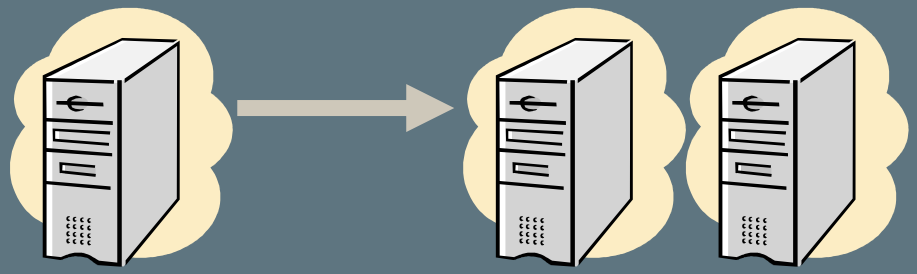


A skálázás módjai



Vertical scaling

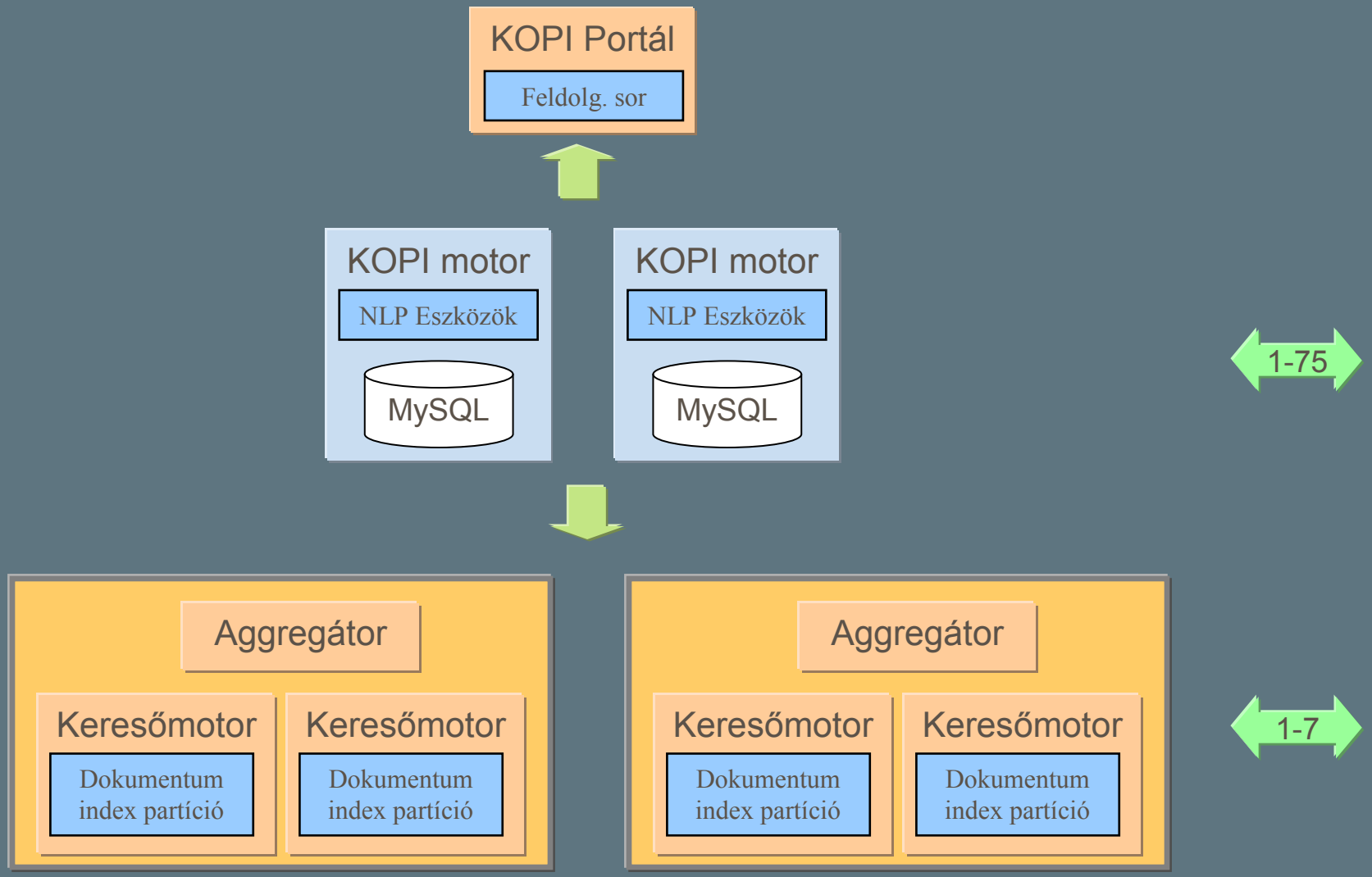
„Scaling up”



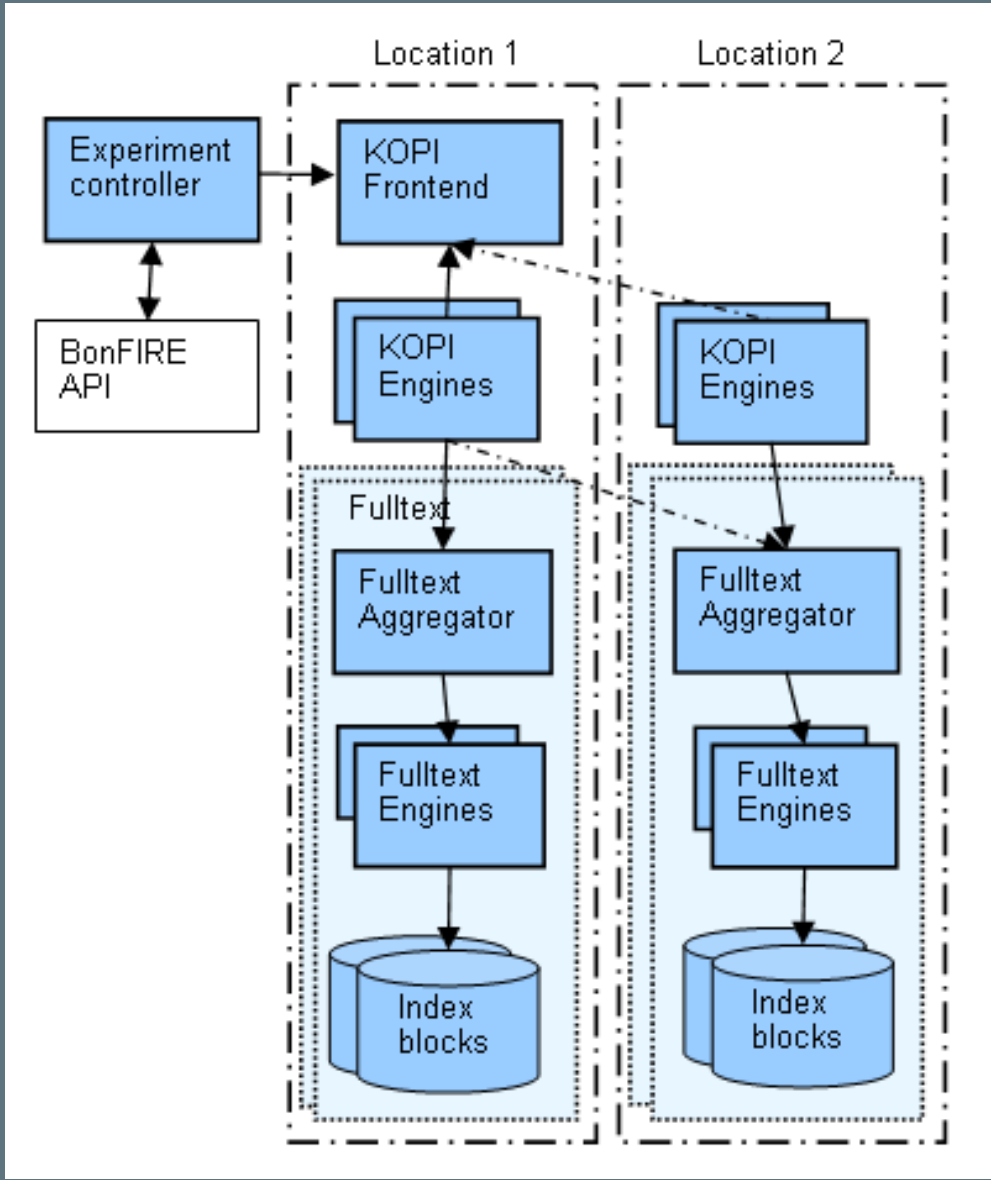
Horizontal scaling

„Scaling out”

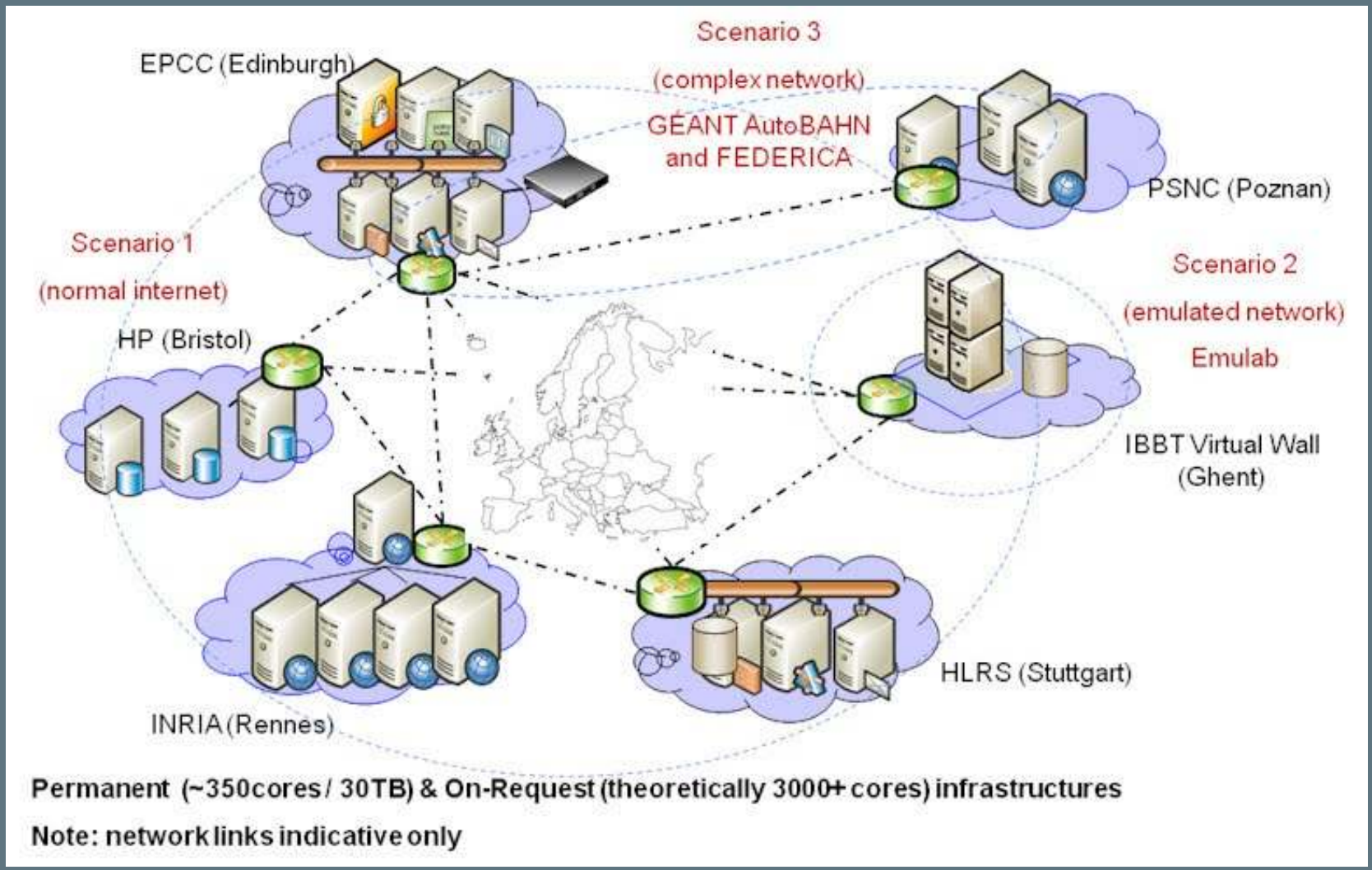
Skálázási lehetőségeink



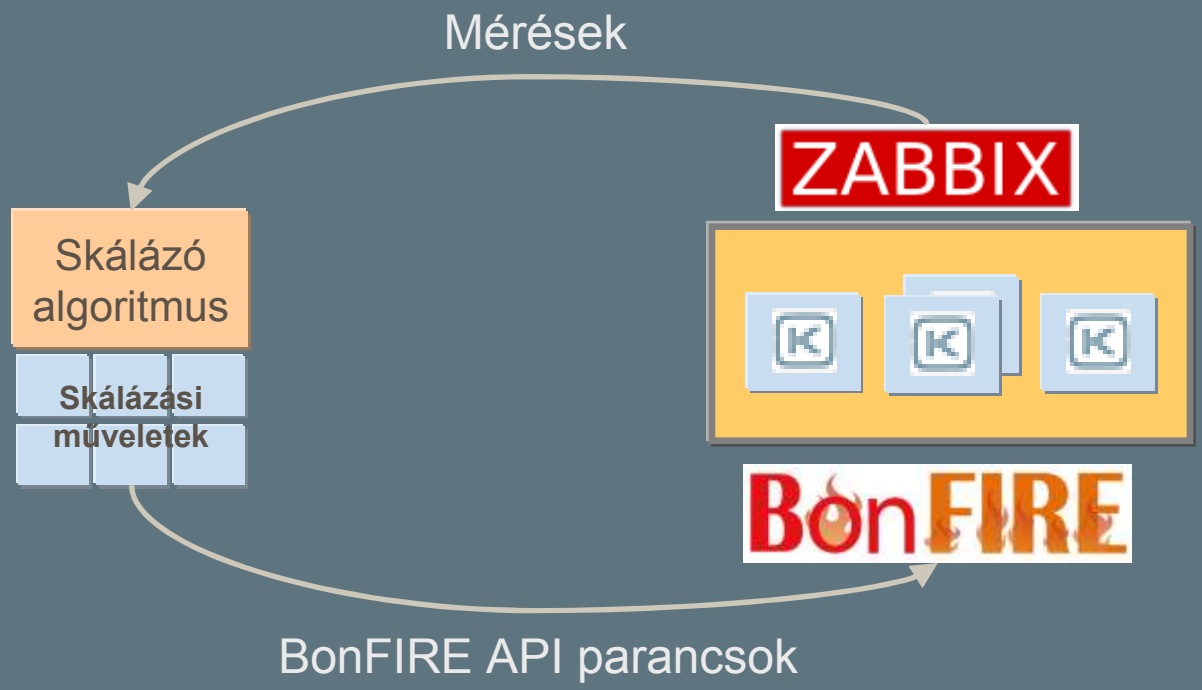
Kísérleti architektúra



BonFIRE: elosztott felhő tesztkörnyezet



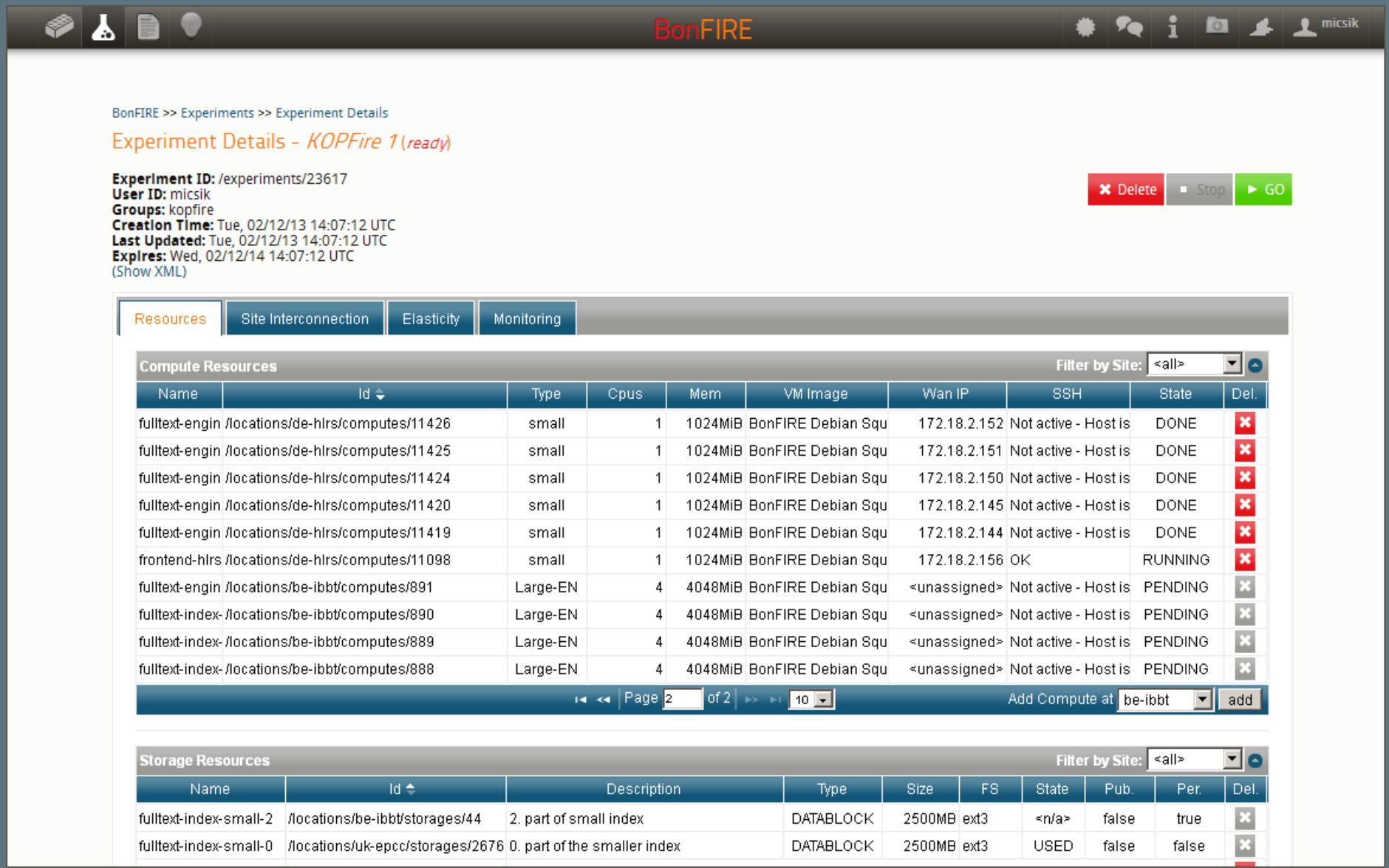
A KOPFire kísérlet



BonFIRE lehetőségek

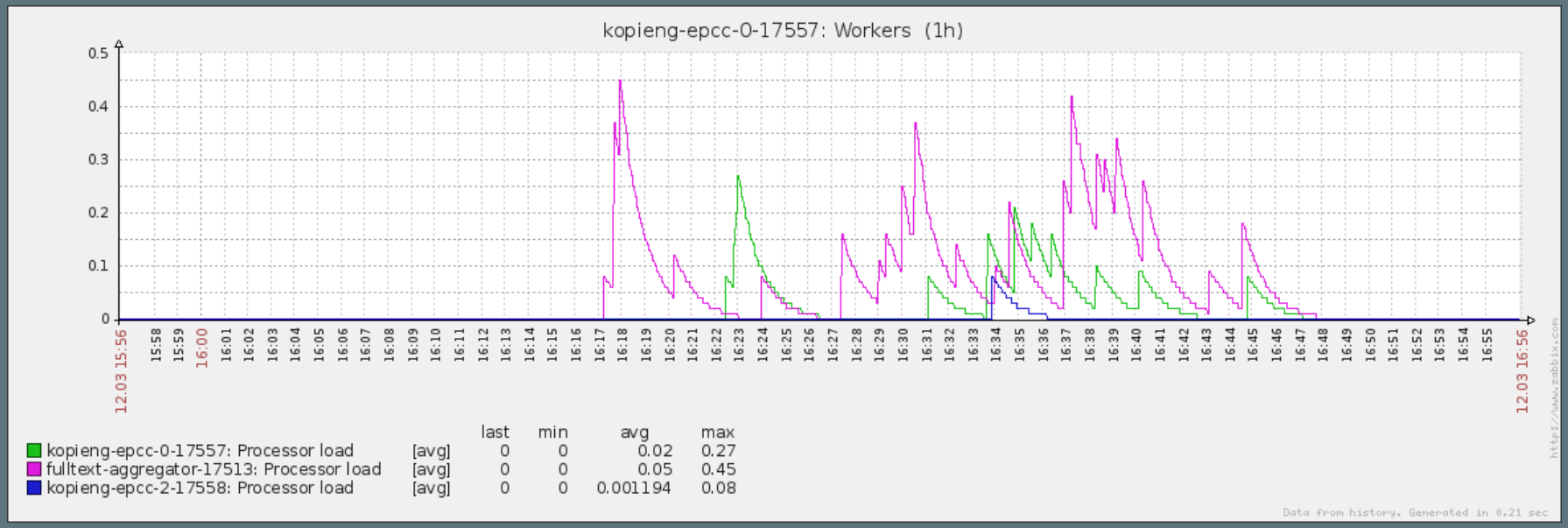
- VM típusok (lite, small, large, stb. + egyedi)
- Adatblokkok
  - OS vagy DATA, perzisztens, shared, stb.
  - Több blokk is kapcsolható egy VM-hez
- Hálózat
  - VM-ek elérhetők: SSH gateway, VPN
  - Speciális lehetőségek: AutoBahn, Virtual Wall, háttérforgalom generálás, stb.
- Monitorozás
- Elasticity as a Service
- Értesítések (RabbitMQ)

A BonFIRE felhasználói portál



Monitorozás

- Egységes monitorozási lehetőség Zabbix-szal
  - Fizikai gépek
  - Virtuális gépek
  - ECO metrics
  - Saját mérések



BonFIRE API

- REST API
  - Experiment descriptor: JSON vagy OCCI

```
<compute xmlns="http://api.bonfire-
    project.eu/doc/schemas/occi">
    <name>my-vm</name>
    <instance_type>lite</instance_type>
    <disk> <storage href="/locations/fr-inria/storages/165"
    /></disk> <nic> <network href="/locations/fr-
    inria/networks/47" /> </nic>
    <location href="/locations/fr-inria" />
</compute>
```

- Restfully (Ruby)

```
experiment.computes.submit(
    :name => "VM#{experiment['id']}",
    :instancetype => "small",
    :disk => [{
          :storage => inria.storages.find{|s|
                s['name'] == SERVER_IMAGE_NAME},
          :type => "OS"    }],
    :location => inria  )
```

- CLI (parancssor)
  - bfcompute create ‘vm0' '/locations/de-hlrs/storages/2088' 23617

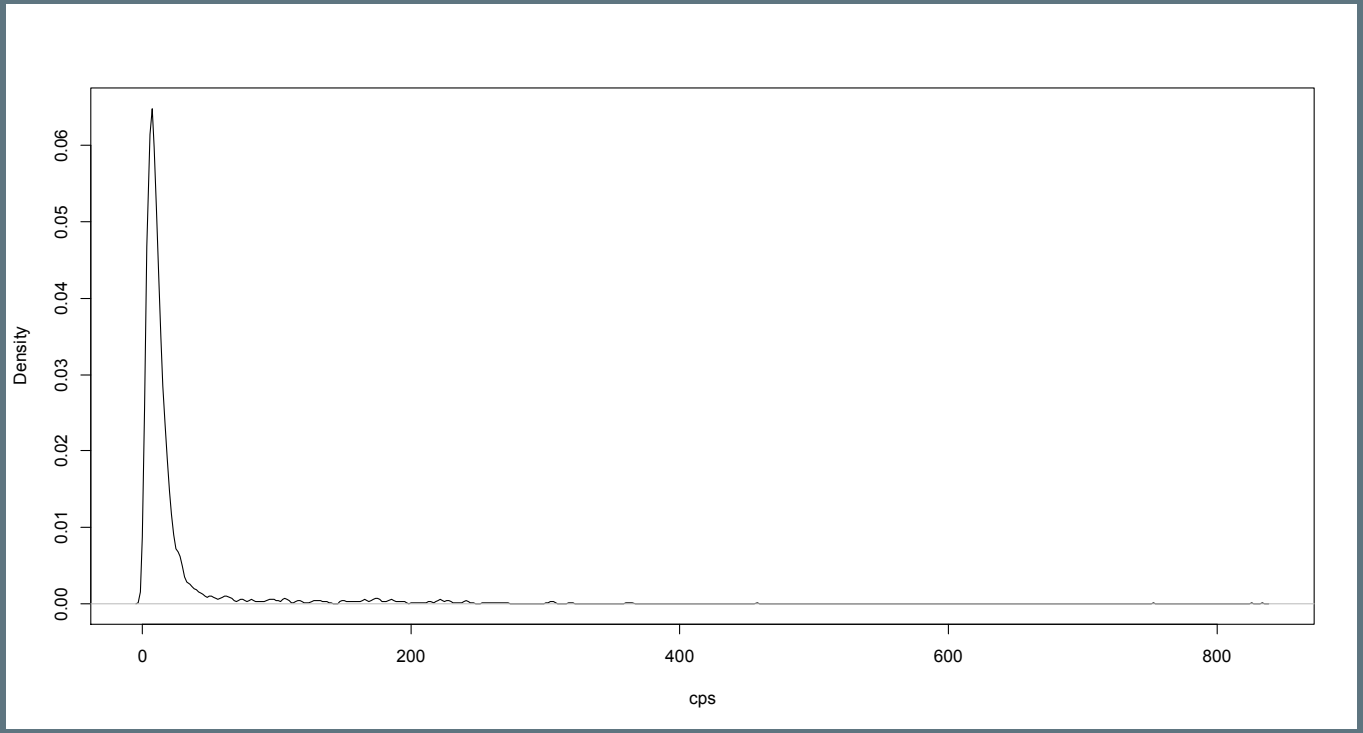
Yours Restfully...

- VM létrehozás
  vm = experiment.computes.submit(...)
- Szoftver futtatás
  vm.waitForState('RUNNING')
  vm.ssh do ….
- Megfigyelés
  values =
  experiment.zabbix.metric('system.cpu.util[,system,avg1]', :type => :numeric, :hosts => vm).values
- Leállítás
  vm.update(:status => 'STOPPED')
  vm.delete

KOPFire rendszerelemek

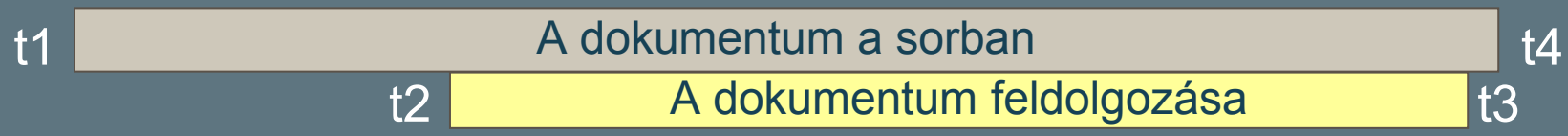
- VM típusok (3 felhőben):
  - (Experiment controller)
  - KOPI Frontend
  - KOPI Engine
  - Fulltext Aggregator
  - Fulltext Engine
- Adatpartíciók:
  - Kis index: 5 x 2.5 GB blocks
  - Nagy index: 11 x 7.2 GB blocks
    - 3 felhőben közvetlenül
    - 4. felhőből NFS-en keresztül

Mit mérjünk?



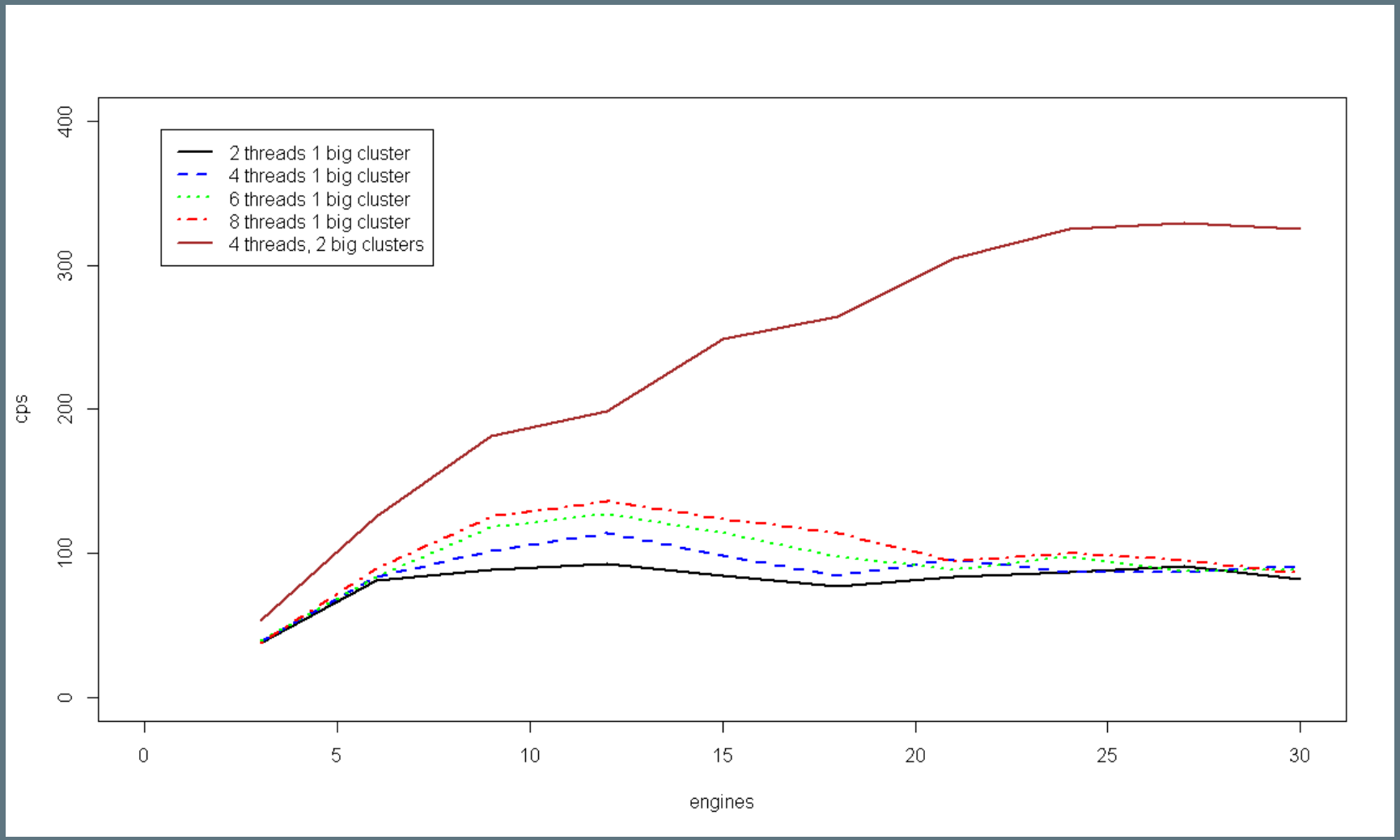
- Kell egy metrika a feldolgozási sebesség mérésére:
  - cps (characters per second)
- Fontos még:
  - Sorban várakozó fájlok és karakterek száma
  - Feldolgozó egységek száma
  - …

Sokféle cps metrika



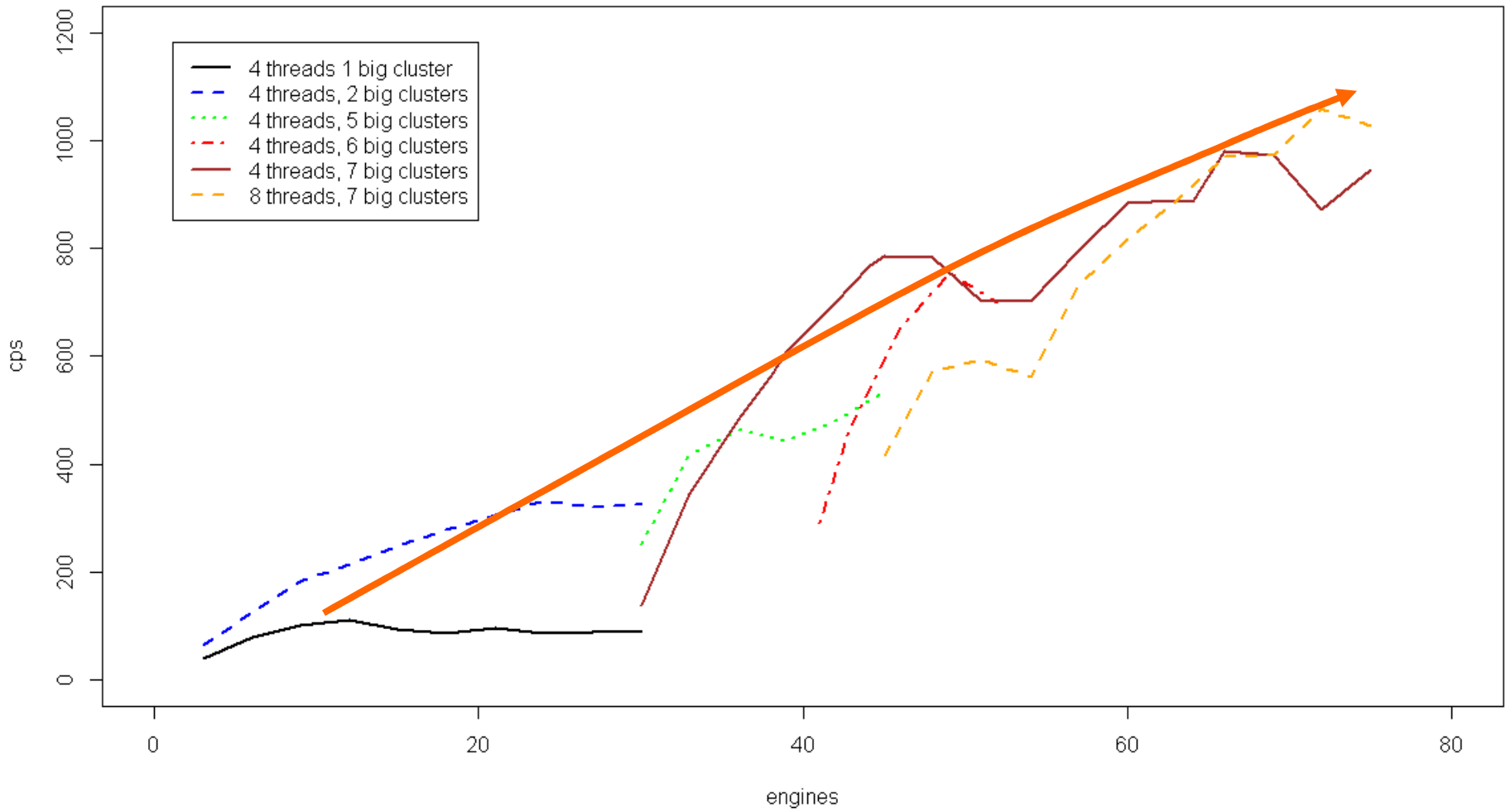
- cps
  - feldolgozott karakterek / utolsó n perc
  - = pillanatnyi feldolgozási sebesség
- pcps (processing cps)
  - dokumentum mérete / (t3 – t2)
- qcps (queue cps)
  - dokumentum mérete / (t4 – t1)
  - = a felhasználó által érzékelt sebesség

Méretezés

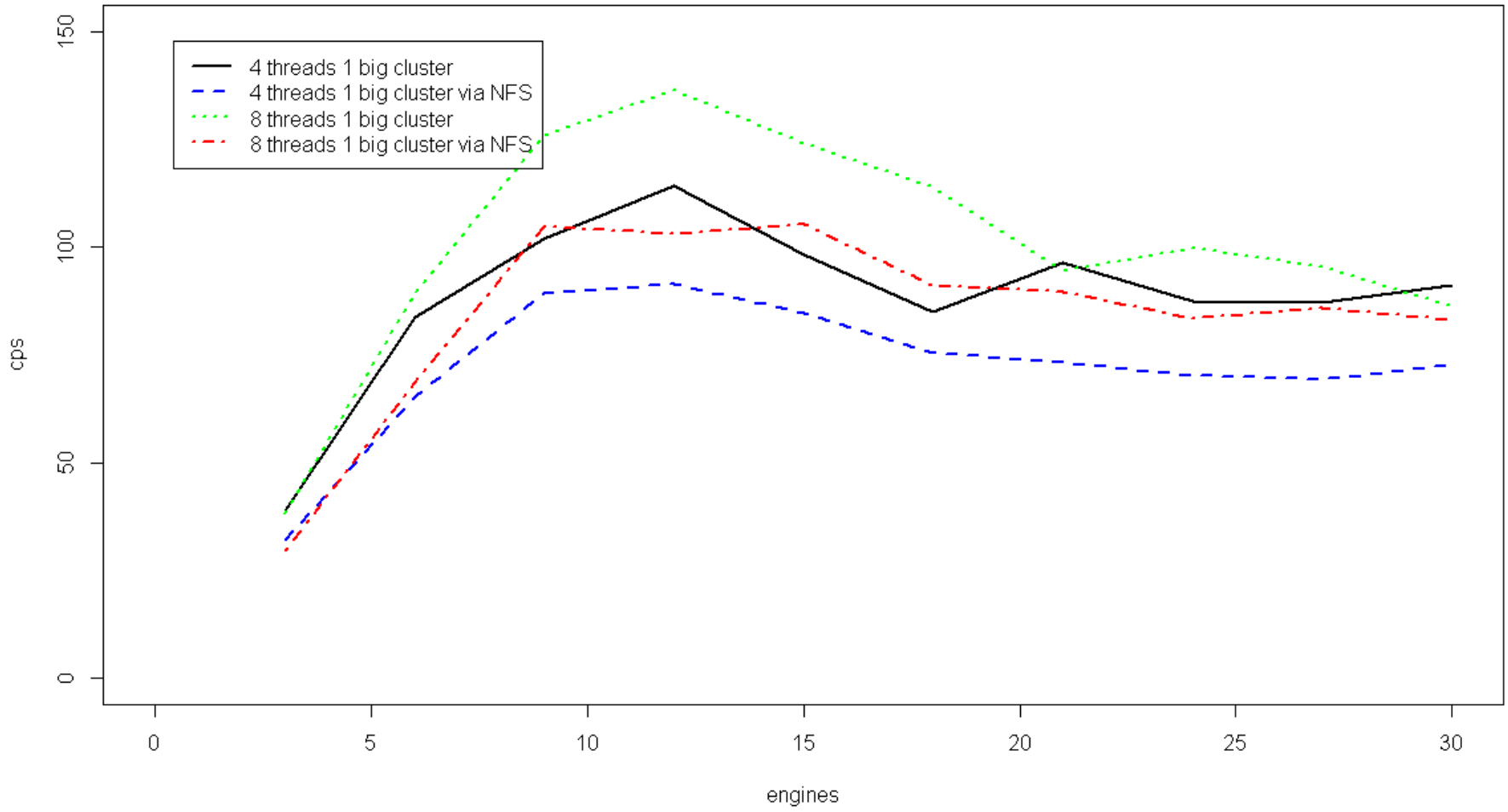


- Mi a rendszerkomponensek optimális aránya?
  - KOPI Engine VM – process-ek – Fulltext Cluster – thread-ek

Teljesítménynövekedés



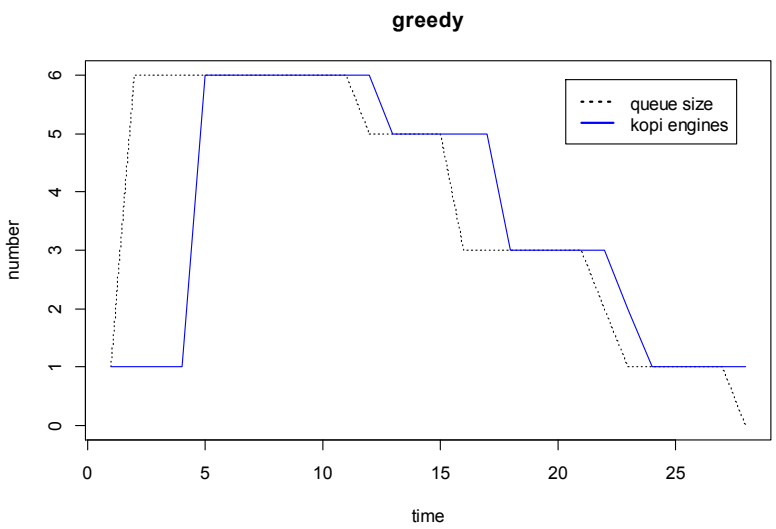
NFS a felhők között?

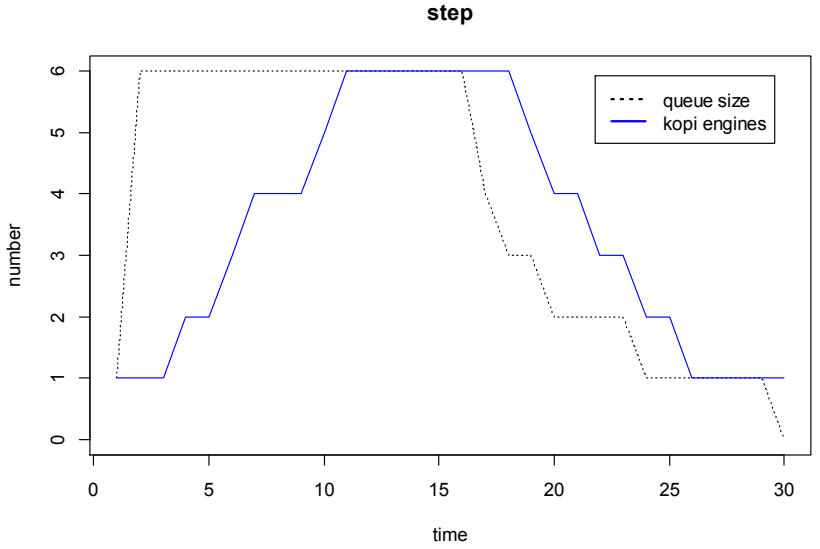


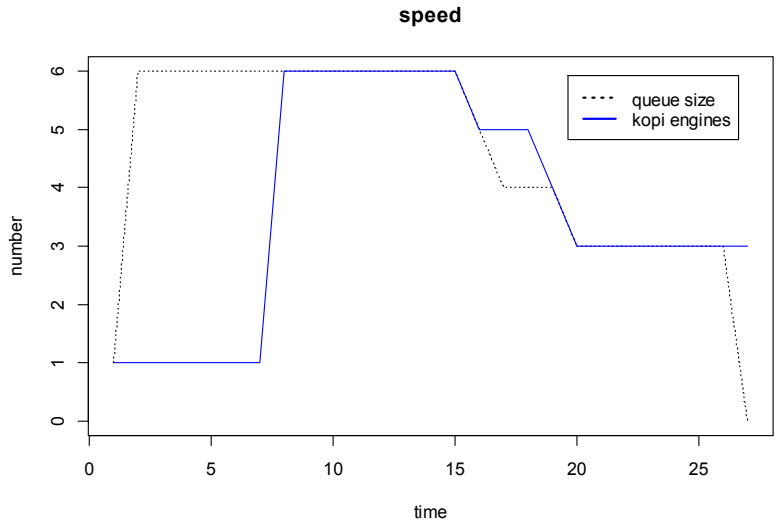
Skálázási lehetőségek

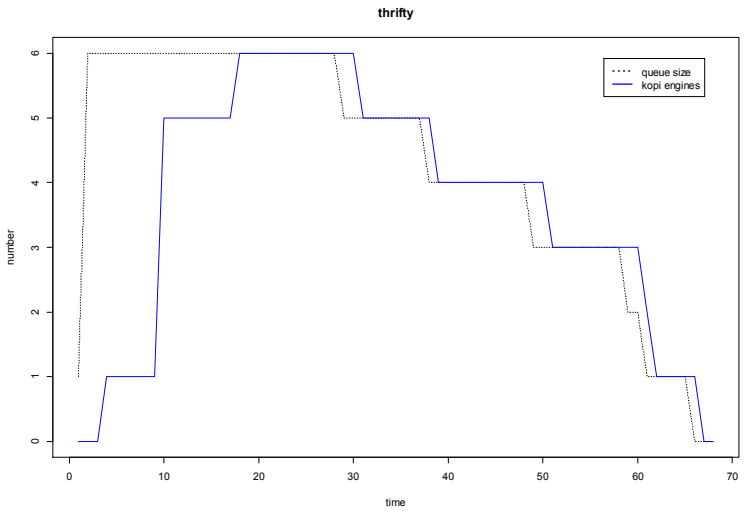
- Különböző skálázási algoritmusokat próbáltunk ki
  - A sorban várakozó dokumentumok száma alapján
    - Kapzsi
    - Takarékos
    - Óvatos
  - Feldolgozási sebesség alapján
    - Tempomat: Adott sebességet közelítő
  - Stb.

Skálázási módszerek









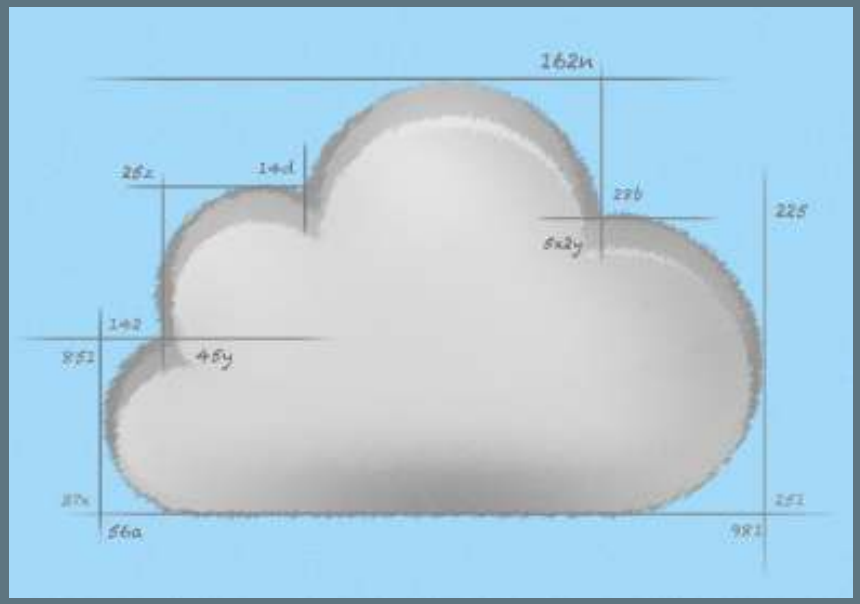
Hibatűrés

- A skálázás során időnként amúgy is rá kell nézni az infrastruktúrára...
- ...ilyenkor szoftver és hardver hibákat is ki lehet küszöbölni...
- ...mivel a komponensek már virtuális gépekben futnak

Szimuláció

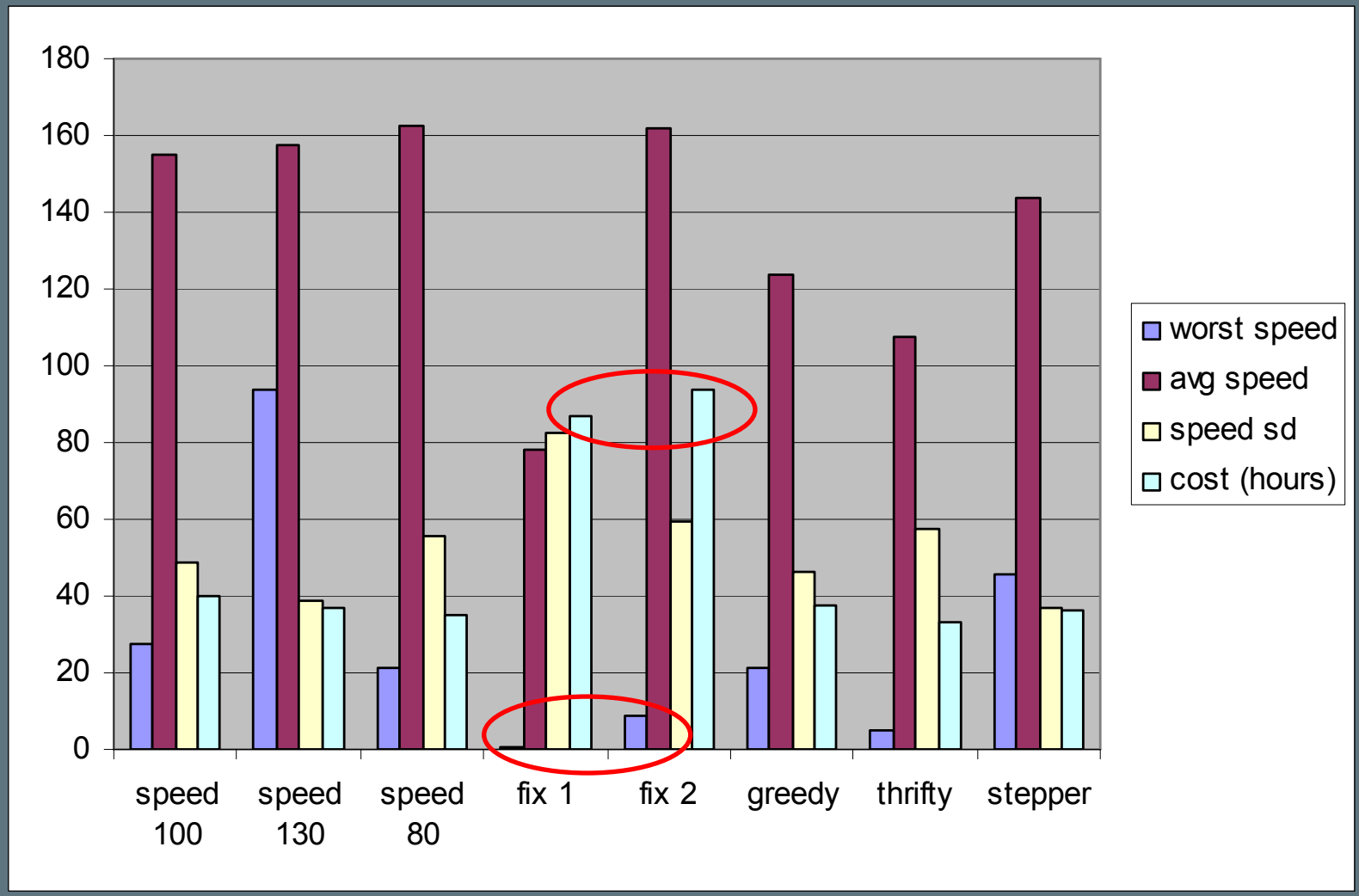
- Pár órás tesztek nem elegendőek
- Hogyan lehetne több havi futáson vizsgálni a skálázódást? -> Szimuláció!
- Az alapadatok ismertek, pl.
  - VM-ek indításához, leállításához szükséges idő
  - Adott VM válaszidejének átlaga, eloszlása

Szimulációk értékelése



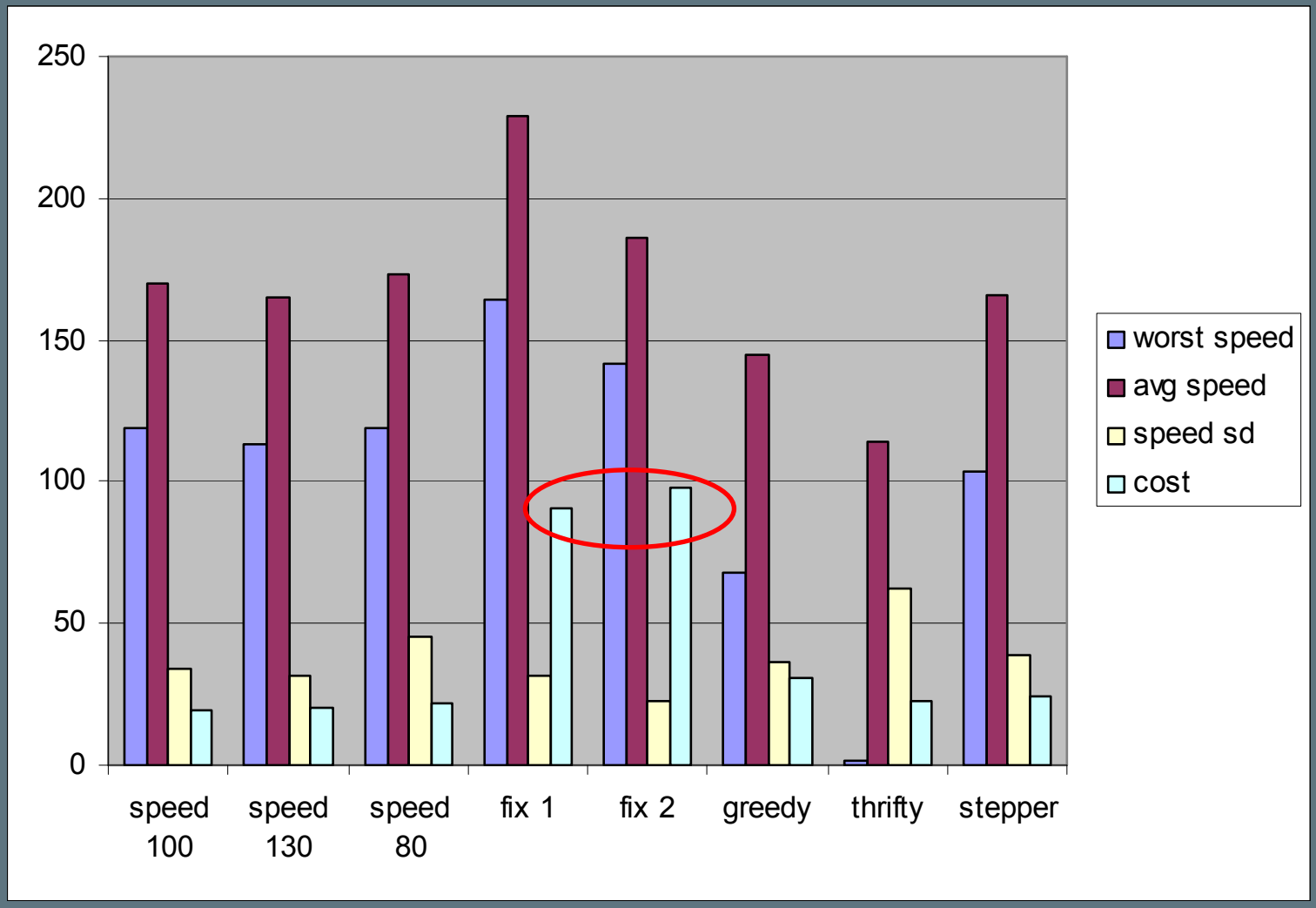
- Mit kell nézni egy hónapos futás esetén?
  - Legrosszabb feldolgozási esetek
  - Átlagos feldolgozási idő
  - A feldolgozási idő szórása
  - A becsült költségek

2012 januári forgalom szimulációja



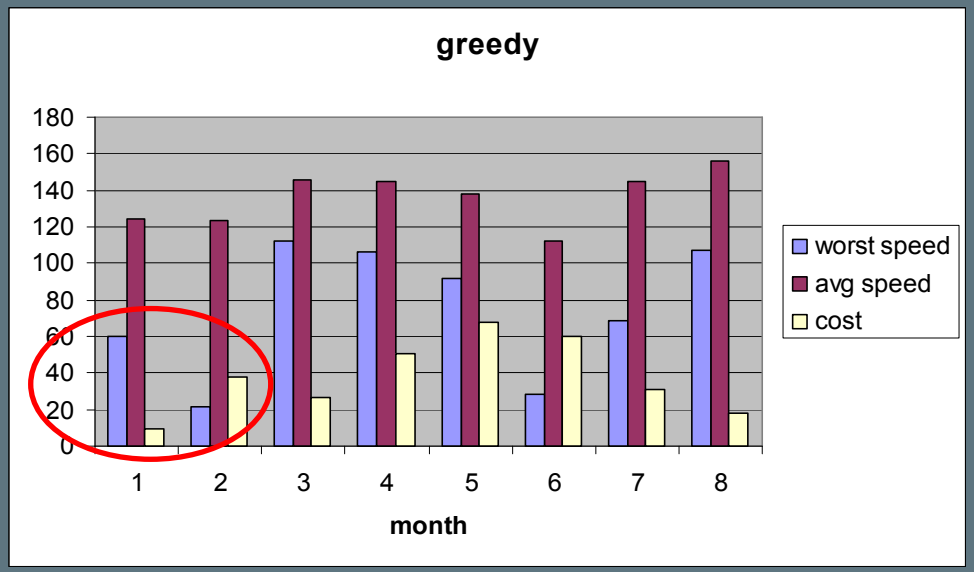
Költségszámítás megkezdett órák alapján

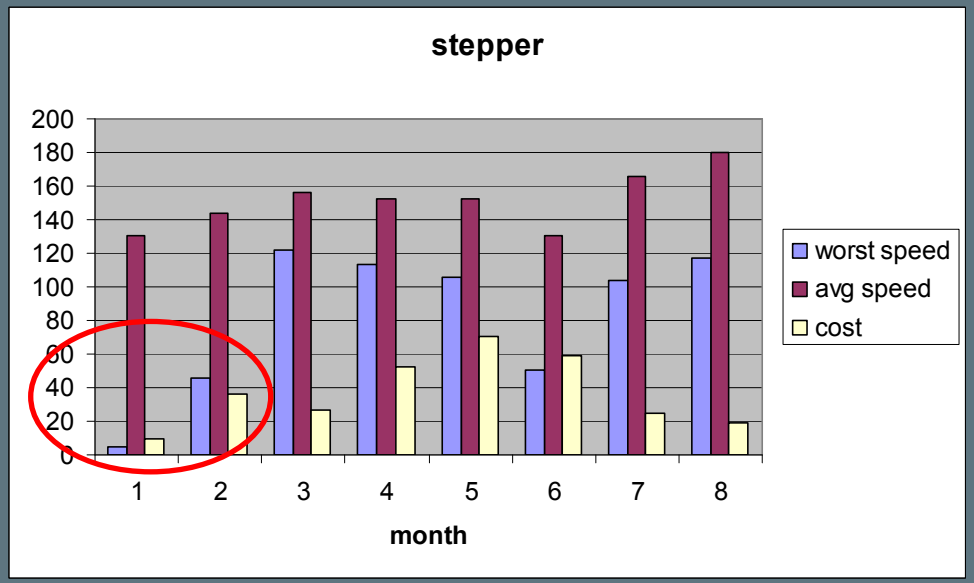
2012 júniusi forgalom szimulációja



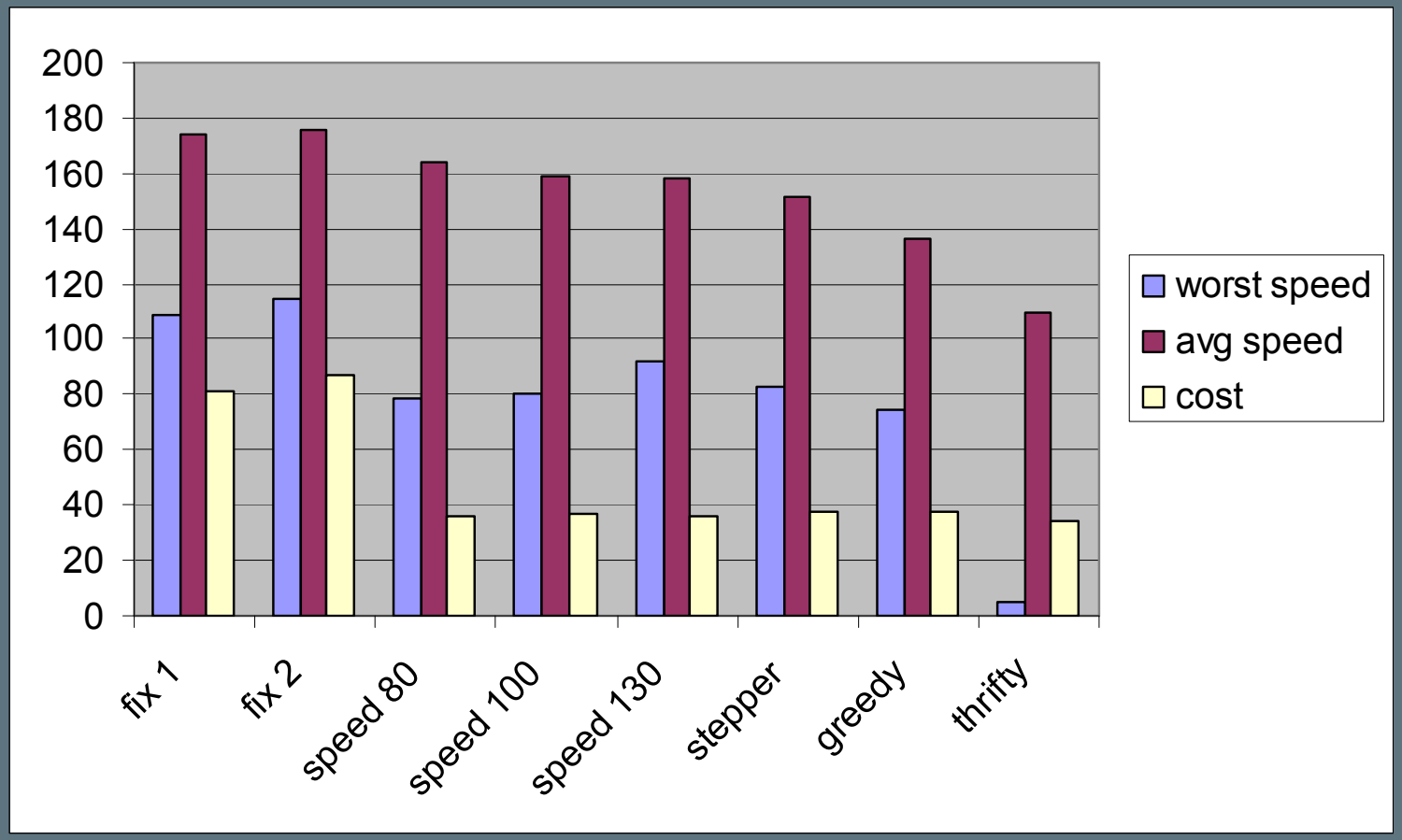
Költségszámítás megkezdett órák alapján

Két módszer összehasonlítása





Végső összehasonlítás



A megjelenített értékek 8 hónap átlagai

Összefoglalás

- Terhelési problémák tesztelésére jó lehetőség egy cloud testbed
- A skálázás nem csak gyorsít, de más előnyökkel is jár, pl. hibatűrés
- A BonFIRE testbed
  - Elérhető lesz legalább 2014 őszéig
  - Nyílt hozzáférés 1 napon belül

www.bonfire-project.eu

kopi.sztaki.hu

Köszönöm a figyelmet!

micsik@sztaki.hu

MTA SZTAKI
Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutatóintézet