Condor
High Throughput Computing

Condor rendszer röviden

Szeberényi Imre

<szebi@iit.bme.hu>

IKTA NI-2000/0008

munkaszkasz zárókonferencia

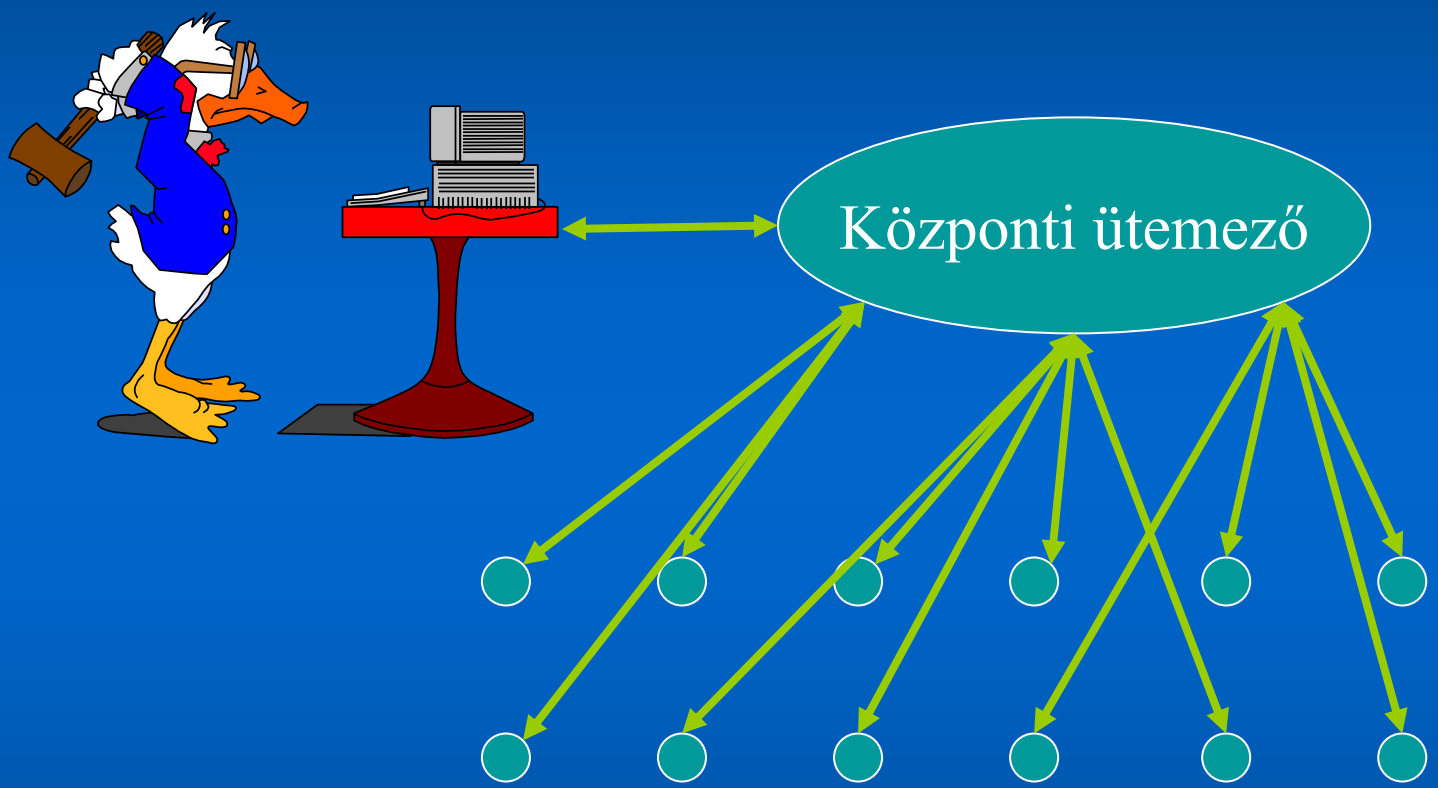
2001.10.12

A Condor rendszer jellemzői

Speciális ütemező (batch) rendszer

- Elosztott, heterogén rendszerben működik.
- Alapvetően a szabad CPU ciklusok kihasználására tervezték.
- Képes egy működő feladatot áthelyezni az egyik gépről a másikra (migráció).
- Az ún. ClassAds mechanizmussal képes a rendszerben levő változó erőforrásokat az igényeknek megfelelően elosztani.

Condor pool



ClassAds lényege

- A rendszerben levő erőforrások különböző jellemzőkkel (teljesítmény, architektúra, op. rendszer, stb.) rendelkeznek.
- A job összeállításánál ezekre a jellemzőkre igényeket lehet előírni, amit a Condor rendszer megpróbál kielégíteni. (Párosítja az igényt az erőforrással)
- A job összeállításánál lehetőség van preferenciák megadására, ami alapján a Condor rangsorolni fog és kiválasztja az igénynek leginkább megfelelő gépet.

Követelmény és rangsor

- Követelmény:

  Requirements = Arch=="SUN4u”

Pontosan kell illeszkednie.

- Rangsor:

  Rank = Memory + Mips

Ha választhat, akkor a nagyobbat fogja választani

A dolgok két oldala (1)

A kifejezések a két hirdetés adatterében értékelődnek ki (adA, adB).

Felhasználó (igénylő) oldala:

```
Requirements = Arch == "INTEL" &&
               OpSys == "LINUX"
Rank = TARGET.Memory * 10 +
       TARGET.Disk + Mips
```

A dolgok két oldala (2)

Erőforrás oldal:

```
Friend  = Owner == "haver"
Trusted = Owner != "judas"
Mygroup = Owner == "zoli" || Owner == "jani"
Requirements  = Trusted && (Mygroup ||
        LoadAvg < 0.5 && KeyboardIdle > 10*60)
Rank = Friend + MyGroup*10
```

Hogyan néz ki egy feladat futtatása ?

- A job összeállítása
- Job bejelentése a Condor-nak
- Job-ot a Condor futtatja az általa kiválasztott gép(eken), szükség esetén átmozgatja egy másik gépre.
- Job befejeződik, a Condor e-mail-t küld a felhasználónak.

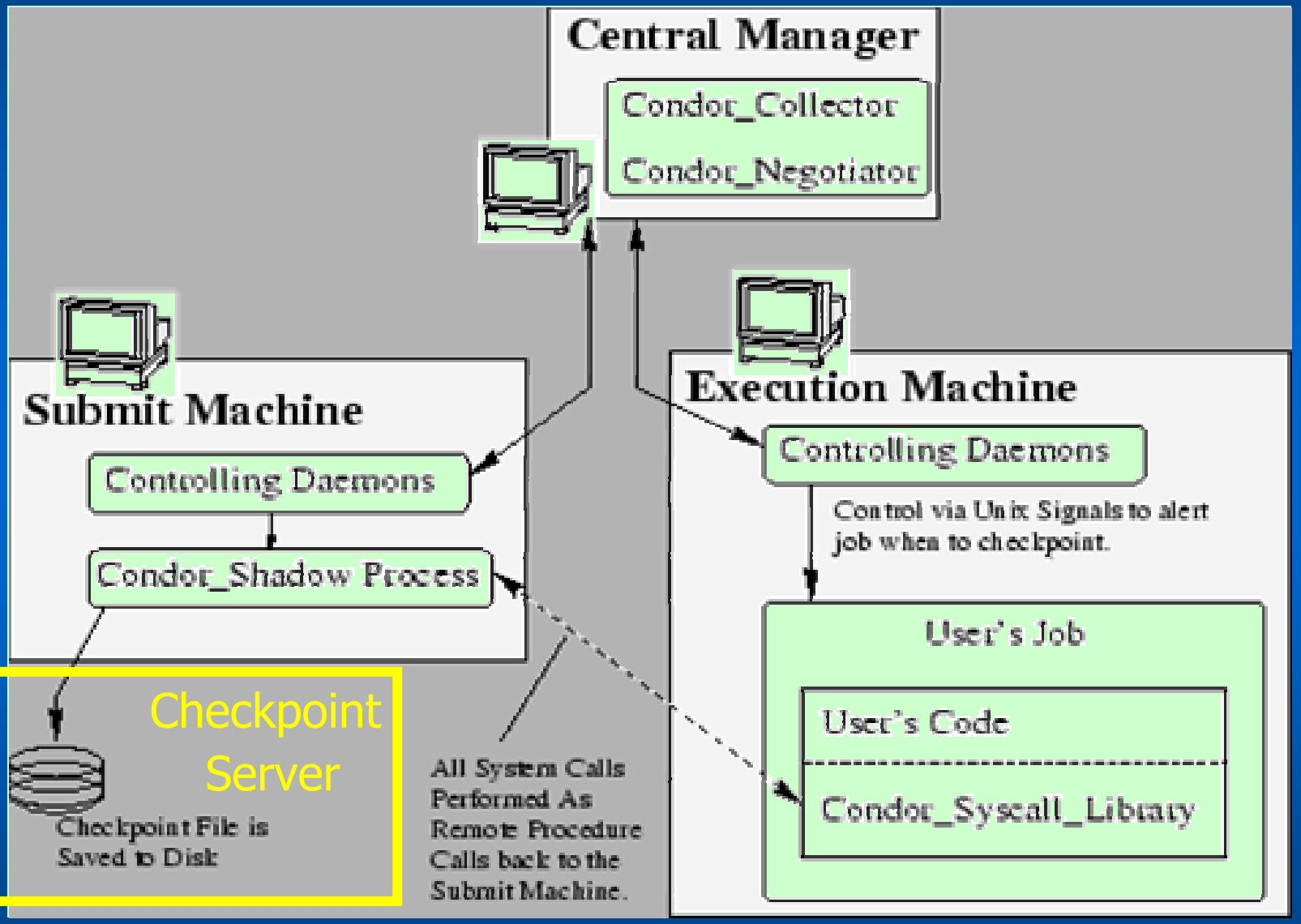
Egy egyszerű jobleíró

```
universe = vanilla
executable = mathematica
input = in$(Process).dat
output = out$(Process).dat
queue 5
```

Feladatkörök

- Central Manager
- Execute Machine
- Submit Machine
- Checkpoint Server

Condor pool elemei



A rendszer részei (1)

- condor_master
  - fő process, ez indítja a többi daemont
- condor_startd
  - execute gépen hirdet
- condor_starter
  - ez indul el először, ha job kerül gépre
- condor_schedd
  - submit queue-t kezeli

A rendszer részei (2)

- condor_shadow
  - a submit gépen fut, a távoli process helyi párja
- condor_collector
  - információ gyűjtő
- condor_negotiator
  - a kérések és az erőforrások párosításáért felelős
- condor_ckpt_server
- condor_kbdd
  - konzol állapotát figyeli (nem minden OS-hez)

A rendszer részei (3)

- Contrib modulok:
  - Condor View
  - Checkpoint server
  - PVM support
  - MPI support
  - Event daemon
  - DAGMan Meta-Sheduler (Directed Acyclic Graph Manager)

Milyen feladatok lehetnek ?

- Elsősorban hosszú futási idejű, számításigényes feladatok.
- Különböző univerzumok léteznek
  - Standard
  - Vanilla
  - PVM
  - MPI
  - Globus
  - Scheduler

Standard univerzum

- checkpointing, automatikus migráció
- meglevő programot újra kell fordítani, esetleg csak linkelni
- az alkalmazás nem használhat bizonyos rendszerhívásokat: pl. fork, socket, alarm, mmap
- („elkapja” a file műveleteket)

Vanilla univerzum

- nincs checkpointing, nincs migráció
- meglevő futtatható kódot nem kell vátoztatni
- nincs korlátozás a rendszerhívásokkal szemben.
- NFS, vagy AFS kell !!!!

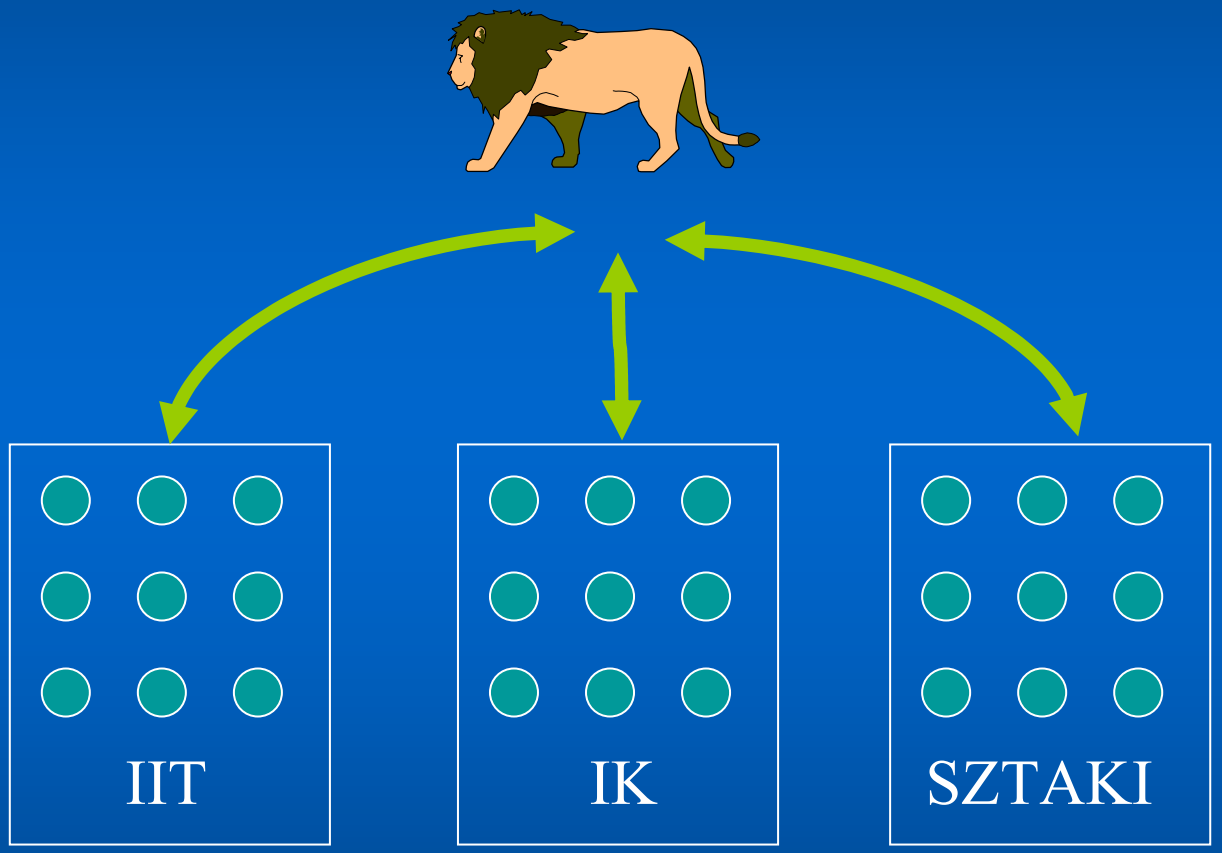
PVM univerzum

- MW jellegű PVM programok környezete
- Binárisan kompatibilis
- PVM 3.4.2 + taszk kezeléshez kieg.
- Dinamikus VM kialakítás.
- Heterogén környezet támogatása
- Egy user csak egy példányban futathat deamont

MPI univerzum

- MPICH változtatás nélkül.
- Bináris kompatibilitás
- Csak ch_p4 device
- Dinamikusan nem változhat
- Nem állhat meg.
- NFS vagy AFS kell.

Barátságos pool-ok (Flock)



Telepítés

- A telepítés bináris disztribúciókból történik.
- Nem minden környezethez létezik binaris disztribúció.
- Elterjedtebb munkaállomások UNIX változatai.
- PC-s környezethez Linux és Solaris van, de pl. FreeBSD nincs.
- NT, de csak Vanilla univerzummal.
- A telepítéshez perl script és kb 350 oldal doksi áll rendelkezésre.

Condor
High Throughput Computing

Egy Condor-PVM alkalmazás és a FLOCK funkció bemutatása

Szeberényi Imre

<szebi@iit.bme.hu>

IKTA NI-2000/0008

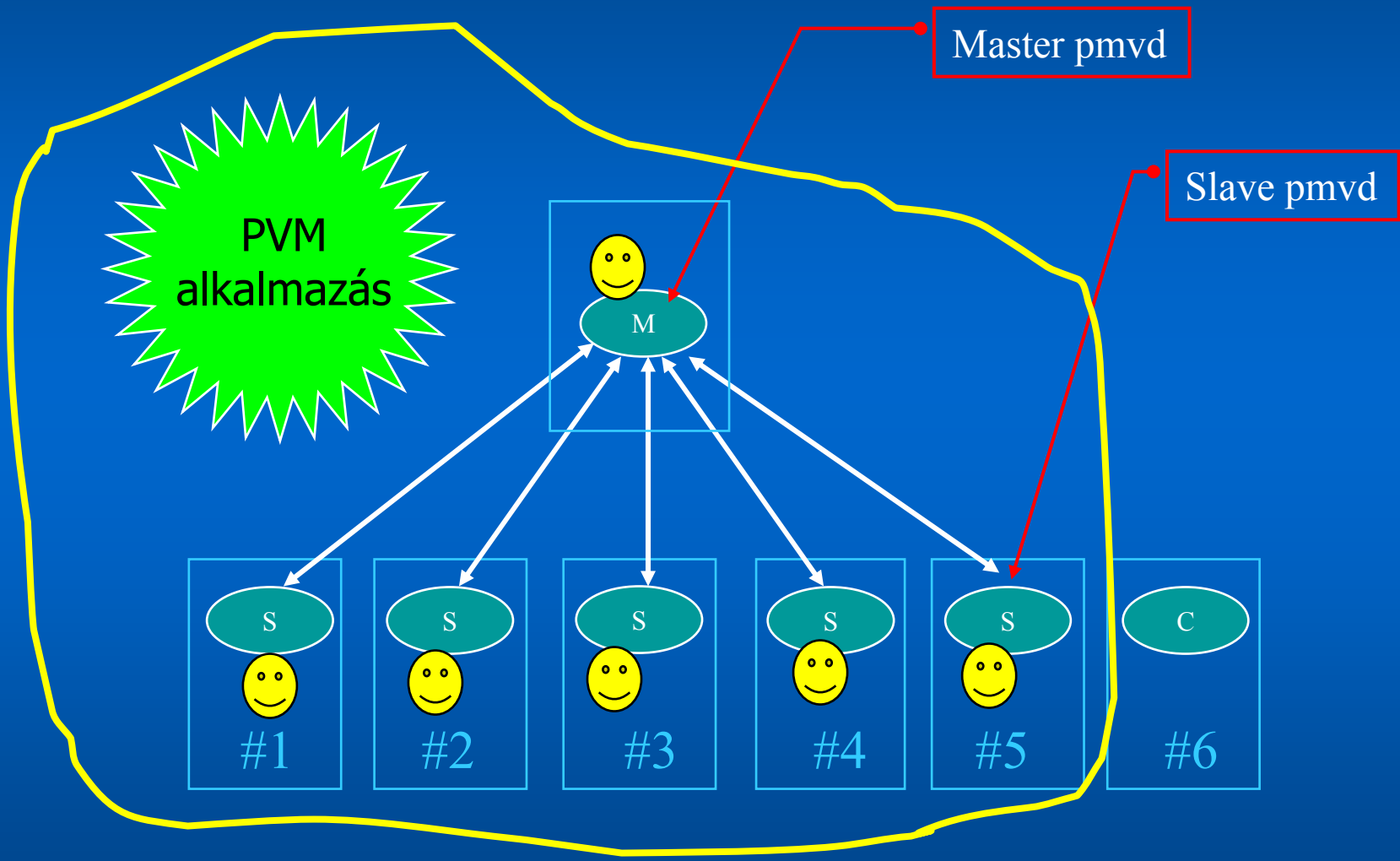
munkaszkasz zárókonferencia

2001.10.12

Condor-PVM

- Nagyobb méretű elosztott PVM rendszer üzemeltetése nem egyszerű feladat.
- Condor-PVM
  - felépíti a virtuális gépet,
  - támogatást nyújt az üzemeltetésében,
  - lehetővé teszi, hogy hibatűrő PVM alkalmazást készítsünk

Condor felépíti a virtuális gépet



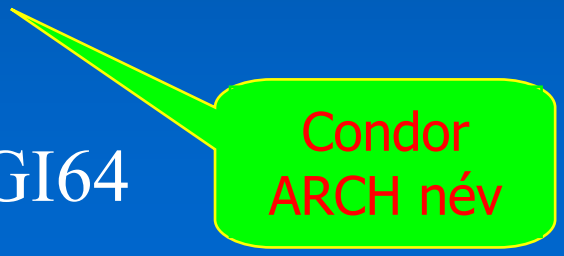
Heterogén PVM program

Szimbolikus linkek alkalmazása:

0 --> SOLARIS26 --> SUN4SOL2

1 --> LINUX

2 --> SGI --> SGI64



Ezek a katalógusok tartalmazzák az architektúra függő részeket. (pl: 0/pi)

pi.cmd jobleíró

```
universe = PVM
executable = SOLARIS26/pi
input = pi.in
output = pi.out
error = pi.err
log = pi.log
```

pi.cmd jobleíró (2)

```
##  Machine class 0 ##
Requirements =
                (Arch == "SUN4u") &&
                (OpSys == "SOLARIS26")
machine_count = 1..3
queue
```

pi.cmd jobleíró (3)

```
##  Machine class 1 ##
Requirements =
                (Arch == "INTEL") &&
                (OpSys == "LINUX")
machine_count = 1..8
queue
```

PVM taszkok indása

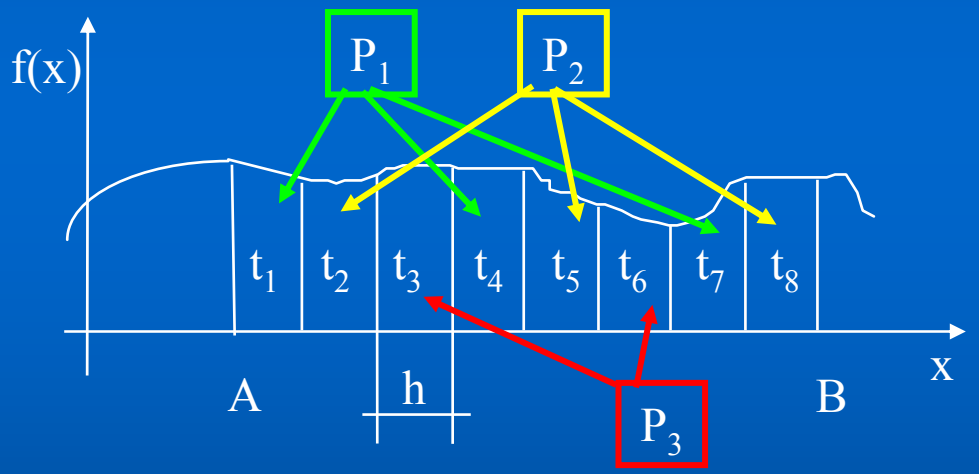
```
pvm_config(&nhost, &narch, &hostinfo);
for (proc = i = 0; i < nhost; i++) {
   arch = hostinfo[i].hi_arch;
   sprintf(taskname, "%s/pi", arch);
   printf("Starting %s\n", taskname); fflush(stdout);
   if (pvm_spawn(taskname, NULL,
      PvmTaskArch, arch, 1, &tids[proc+1]) == 1) {
      proc++;
      if (proc >= nprocs) break;
   }
}
```

Demo: Párhuzamos számítási példa

Számítsuk ki az $\int_A^B f(x)\,dx$ integrál értékét egyszerű numerikus közelítéssel!

$$\int_A^B f(x)\,dx = h \cdot \sum_{i=1}^{N} f\left(A - \frac{h}{2} + i \cdot h\right) \qquad \text{ahol } h = \frac{B - A}{N}$$

N=8 esetén pl:



Az egyes téglányok számítása egymástól függetlenül, párhuzamosan is elvégezhető. Pl. minden task csak minden M-edik téglányt számol ki, majd az összegezzük az eredményeket.

PVM alk. további processzeket hoz létre

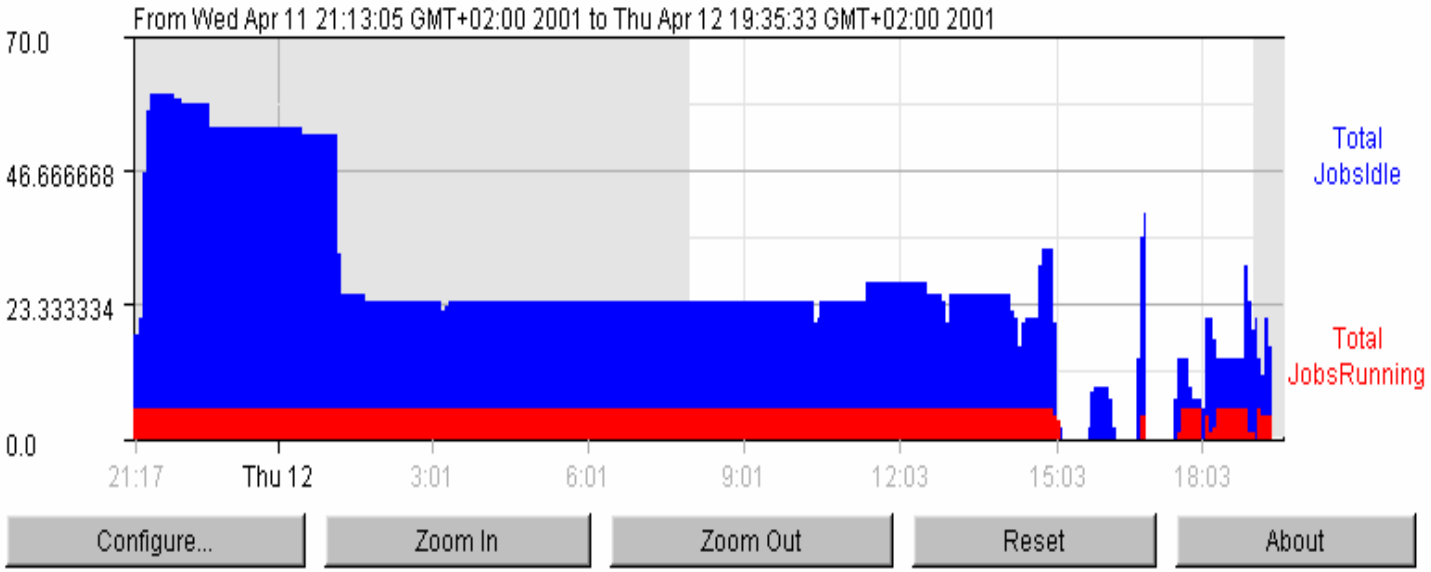
Nhost=13, Narch=5
Starting 0/pi, arch= 0(0), proc=1
Starting 0/pi, arch= 0(1), proc=2
Starting 0/pi, arch= 0(2), proc=3
Starting 1/pi, arch= 1(3), proc=4
Starting 1/pi, arch= 1(4), proc=5
Starting 1/pi, arch= 1(5), proc=6
Starting 1/pi, arch= 1(6), proc=7
Starting 1/pi, arch= 1(7), proc=8
Starting 1/pi, arch= 1(8), proc=9
Starting 1/pi, arch= 1(9), proc=10
Starting 1/pi, arch= 1(10), proc=11
Starting 4/pi, arch= 4(11), proc=12
Starting 4/pi, arch= 4(12), proc=13

Fut a PVM alkalmazás

First instance on bagira: x = 0.22647
First isntance got x = 0.22590 from bvp6
First isntance got x = 0.22575 from bvp2
First isntance got x = 0.22361 from bvp7
First isntance got x = 0.22333 from bvp7
First isntance got x = 0.22304 from bvp8
First isntance got x = 0.22347 from bvp2
First isntance got x = 0.22319 from bvp3
First isntance got x = 0.22290 from bvp3
First isntance got x = 0.22633 from bagira
First isntance got x = 0.22276 from kempelen
First isntance got x = 0.22619 from bagira
First isntance got x = 0.22604 from bagira
First isntance got x = 0.22261 from kempelen
sum = 3.14159298 err = 2.384186e-07

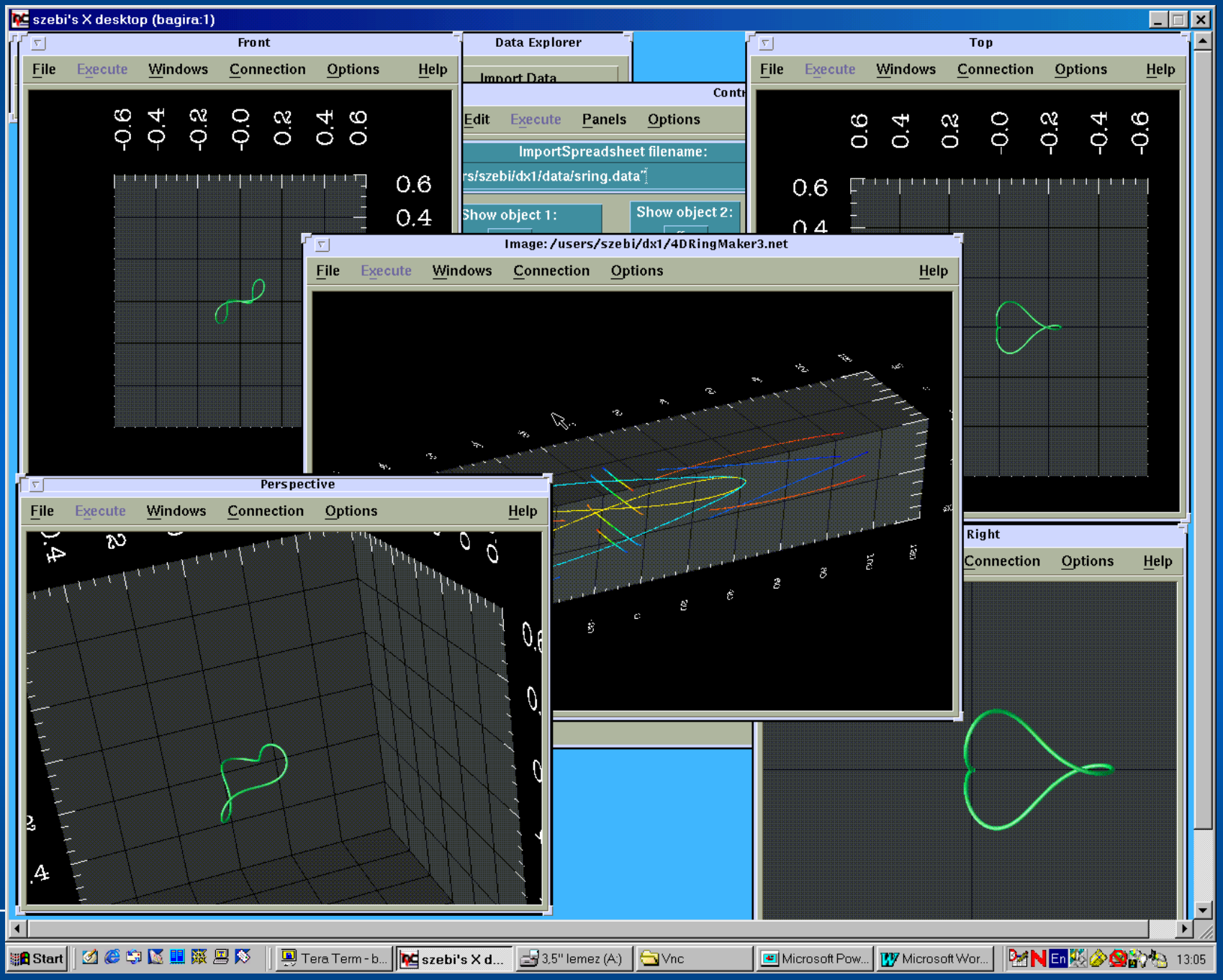
Egy nap statisztikája

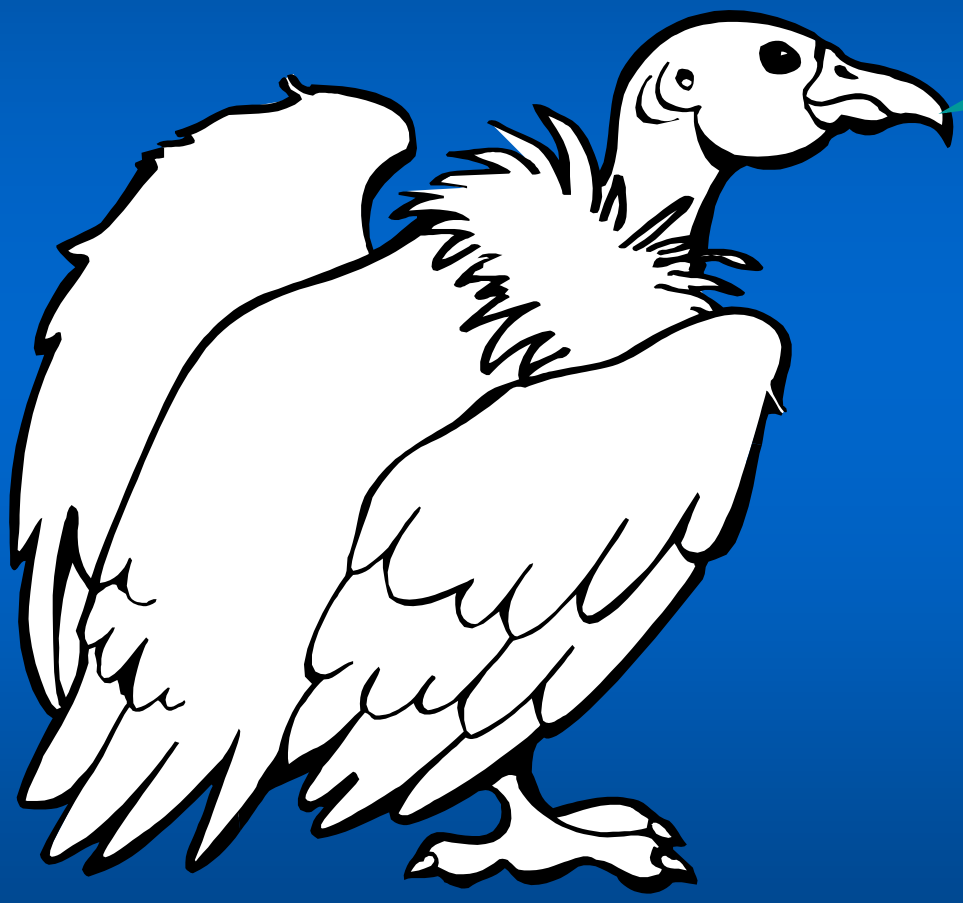
TU-Budapest IIT Condor Pool User Statistics for Day



[Graph Hints: The Y-axis is number of jobs, the X-axis is time. When graph finishes updating, press "Configure.." to view different Architecture or State data. Also, you can use the mouse to draw a rectangle on the graph and then press "Zoom In". Press "Reset" to center/resize the data after Configure or when done zooming. Nighttime shows up on graph background as grey.]

User	Total Allocation Time (Hours)	JobsRunning Average	JobsIdle Average	JobsRunning Peak	JobsIdle Peak
Total	95.8	4.7 (19.0%)	23.9 (81.0%)	5.0 (100.0%)	55.0 (100.0%)


szebi's X desktop (bagira:1)
Front
File Execute Windows Connection Options Help
Data Explorer
Import Data
Edit Execute Panels Options
ImportSpreadsheet filename:
rs/szebi/dx1/data/sring.data
Show object 1:
Show object 2:
Top
File Execute Windows Connection Options Help
Image: /users/szebi/dx1/4DRingMaker3.net
File Execute Windows Connection Options Help
Perspective
File Execute Windows Connection Options Help
Right
Connection Options Help
Start
Tera Term - b...
szebi's X d...
3,5" lemez (A:)
Vnc
Microsoft Pow...
Microsoft Wor...
13:05



Köszönöm a figyelmet!