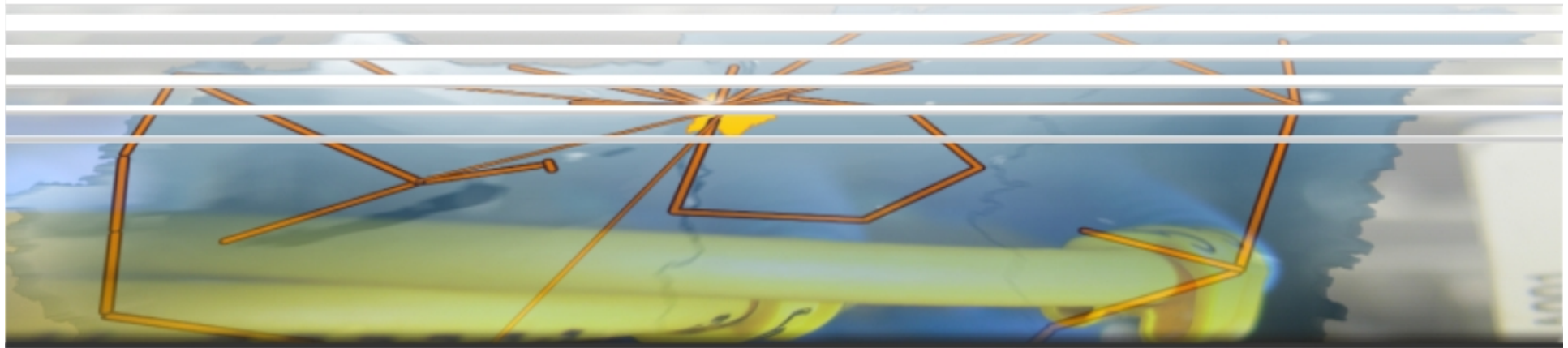


Szuperszámítógép Felhasználói Tájékoztató



11/18/11
Debrecen

Stefán Péter
<szuper@niif.hu>



Miről lesz szó?

- Rövid történeti áttekintés.
- Az NIIFI szuperszámítógépei, jellemzőik.
- Hozzáférés, felhasználói kör.
- Használat:
 - belépés,
 - fordítás, párhuzamos feladatok (MPI, PVM, OpenMP) készítése,
 - kötegelt feladatok futtatása,
 - párhuzamos feladatok futtatása,
 - Eredmények kinyerése.
- Kitekintés a nagyvilágba, gyártók, trendek.
- Összefoglalás, trendek.

I. Bevezetés, áttekintés

Történeti áttekintés

- Elosztott számítástechnika, nagy teljesítményű számátástechnika.
- A számítástechnika egy speciális területe, amely formalizált **tudományos modellek** futtatására szolgál (numerikus analízis).
- A számítástechnika bölcsőjétől jelen van. Neumann János differenciál egyenleteket szeretett volna kora tudósaihoz képest hatékonyan megoldani.
- Átfogó terület: **a hardvertől az alkalmazásokig** a tématerület teljes egészét átöleli.



Történeti áttekintés

- Számítási erőművek, amelyek milliárd egész és lebegőpontos matematikai műveletet képesek elvégezni egy másodperc alatt.

$$\frac{1.678999 \cdot 10^5}{2.99765 \cdot 10^9}$$

- Floating point per second: flop/s. **Tflops** - 10^{12} . **Pflops** - 10^{15} .
- Megkülönböztető jellemzőjük minden korban:
 - nagyszámú **párhuzamos feldolgozóegység** (processzor);
 - relatíve **nagy memória**;
 - ezek **gyors** összeköttetése.

Történeti áttekintés

- Jelenlegi megoldások:
 - klasszikus **SMP/NUMA** rendszerek (korábban mainframe-ek);
 - **fürtözött** rendszerek (személyi számítógép cluster).
- E két technológia valaha nagyon különbözött, ma azonban azt látjuk, hogy ezek folyamatosan simulnak egymásba: mindkettőben megjelennek a legfontosabb technológiai elemek.
- Például: domain-ek kezelése, PCI busz-technológiák.

NIFI szuperszámítógépek

- 11 éve foglalkozunk elosztott számítási és adattárolási környezetek kialakításával és működtetésével:
 - 2001 – SUN Fire 10K rendszer (top500 428, hely, 60Gflops, SMP, 96 proci, 48GB mem);
 - 2002 – grid infrastruktúra projektek;
 - 2005-2009 – E15K upgrade (több körben, 600Gflops, NUMA, 216 mag, 400Gb mem);
 - 2009 – HBONE+ projekt előkészítése;
 - 2010-től – szuperszámítógépek beszerzése, telepítése, üzembe helyezése;
- Kihasználtság: 100%.

NIIFI szuperszámítógépek

- A fejlesztés főbb mérföldkövei:
 - 2009 – előkészítés;
 - 2010 – tendereljárás (sajnos hosszú!!!);
 - 2011 – szállítás, telepítés, tesztelés, átadás.
- Négy helyszín:
 - Debreceni Egyetem;
 - Pécsi Tudományegyetem;
 - Szegedi Tudományegyetem;
 - NIIFI.
- Minden helyszínről delegált szakértői stáb.
(Debrecen: dr. Gál Zoltán.)
- Forrás: ÚMFT HBONE+.

Tervezési szempontok

- Elosztott környezet, **több helyszín** (vs. központilag elhelyezett erőforrás).
- Alapvetően **heterogén platform**: NUMA + MPP + cluster.
- Környezetbarát, **előremutató** technológiák.
- Bővíthetőség.
- Legalább **100-szoros** teljesítménynövekedés.
- Külön-külön is nagyszámítógépek.
- Linux + **OpenSource** technológiák.

Debreceni Egyetem – Altrix ICE

- SGI Altix ICE 8400EX
- Cluster architektúra
- Intel Xeon (Westmere-EP) processzorok
 - 18 Tflop/s (3000 művelet/fő/másodperc);
 - 1536 db core (3.33 GHz);
 - Redundáns Infiniband QDR interconnect;
 - 6 TB memória;
 - ~500 TB háttértár;
 - SuSE Linux operációs rendszer;
 - Vízhűtéses rackek!
 - Nvidia Quadro FX5800 alapú vizualizációs alrendszer (szerver).



Pécsi Tudományegyetem – Ultra Viloet

- SGI UltraViolet 1000 (SGI UV)
- ccNUMA (**SMP**) architektúra!
- Intel Xeon X7542 (Nehalem EX) processzorok
 - 10,5 Tflop/s;
 - 1152 db core;
 - Numalink5 interconnect;
 - 6 TB memória;
 - ~500 TB háttértár;
 - SuSE Linux operációs rendszer;
 - Vízhűtéses rackek!
 - Nvidia Quadro FX5800 alapú vizualizációs alrendszer (szerver).



Szegedi Tudományegyetem – HP Blade

- Hewlett-Packard CP4000BL
- Fat-node cluster architektúra! (blade)
- AMD Opteron 6174 (Magny Cours) processzorok (12 mag/processzor)
 - 14 Tflop/s;
 - 2112 db core;
 - 48 core/node! (SMP jelleg...)
 - 5,6 TB memória;
 - Redundáns QDR Infiniband mesh interconnect;
 - ~250 TB háttértár;
 - RedHat Linux operációs rendszer;
 - Nvidia Quadro FX5800 alapú vizualizációs alrendszer (szerver).

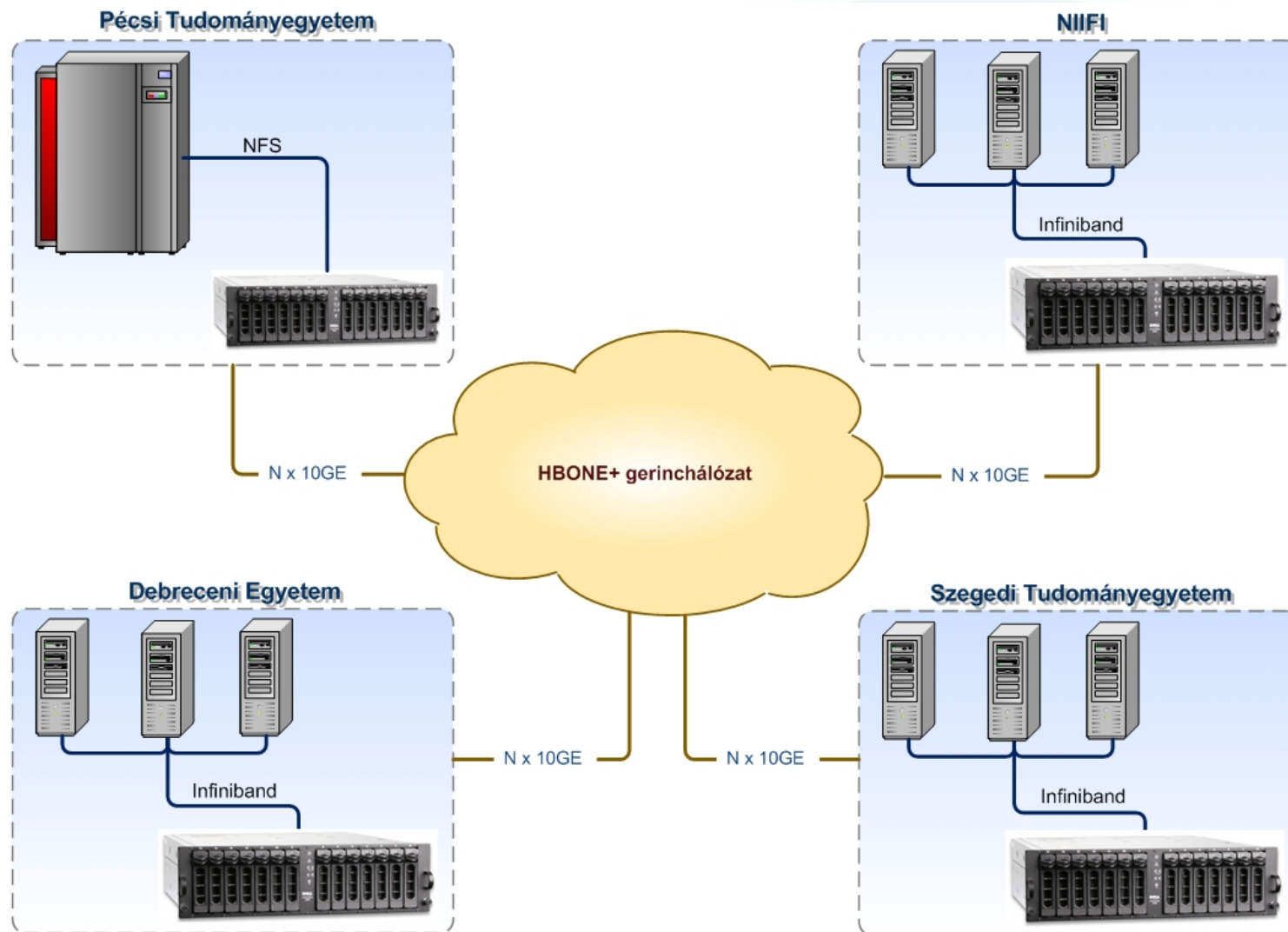


NIIF központ, Budapest – HP pizzadoboz

- Hewlett-Packard CP4000SL
- Fat-node cluster architektúra
- AMD Opteron 6174 (Magny Cours) processzorok (12 mag/processzor)
 - 5 Tflop/s;
 - 768 db core;
 - 24 core/node;
 - 2 TB memória;
 - Redundáns QDR Infiniband mesh interconnect;
 - ~50 TB háttértár;
 - Zárt rendszerű, vízűtési rack!
 - Redhat Linux operációs rendszer.



NIIFI szuperszámítógép projektje



II. Felhasználók, hozzáférés

Területek, ahol használhatók

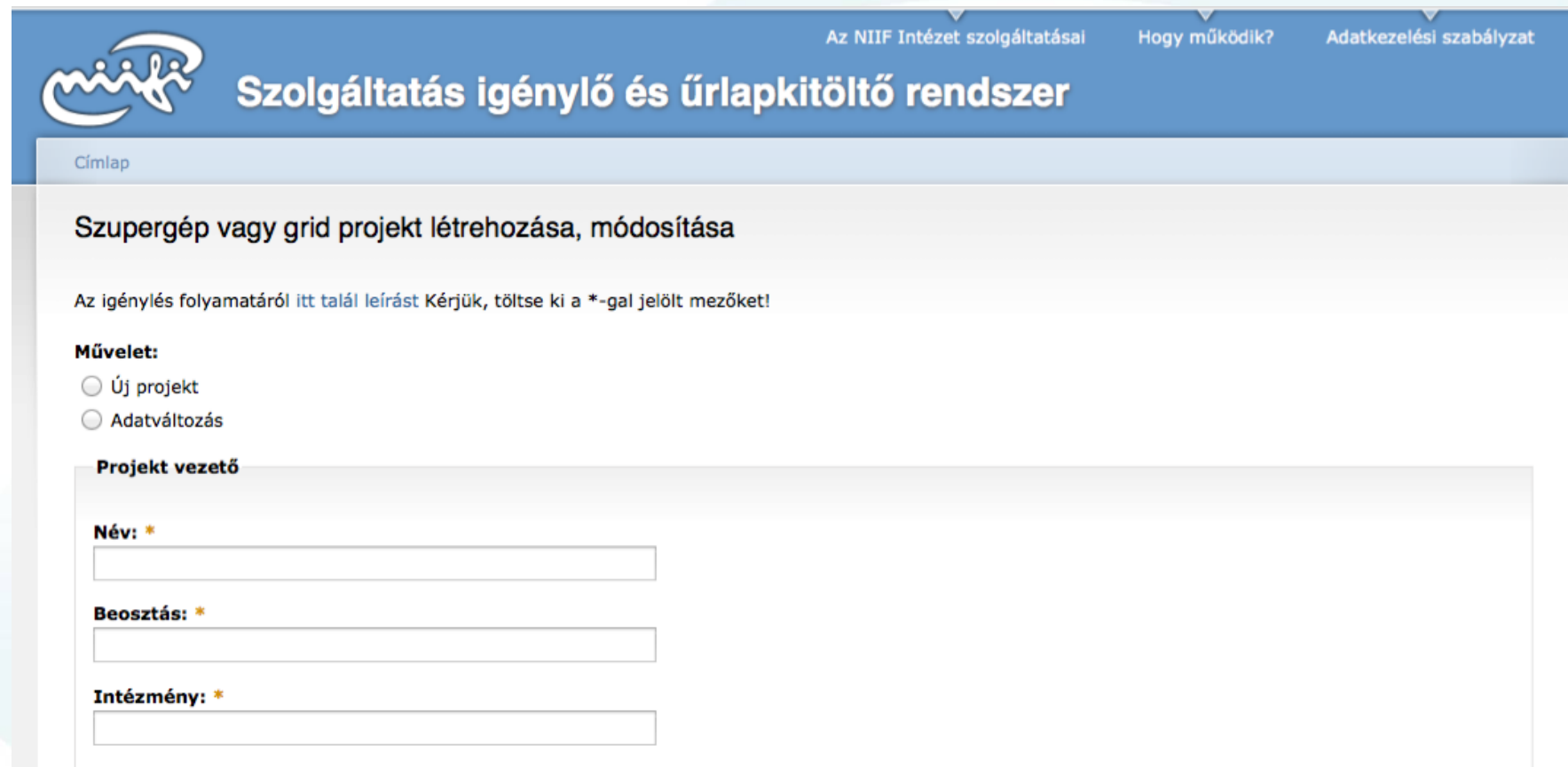
- Mind a **tudományos** mind a cégekre jellemző innovatív területeken használják.
- Csillagászat, bioinformatika (rákkutatás, adatfelismerés), fizika, kémia, matematika, számítástechnika.
- Ipari környezetben: rajzfilmgyártás, autógyártás (töréstesztek, aerodinamikai szimulációk), energetika, pénzügyi elemzések.
- Mindenütt, ahol a kor mércéjével mérve **extrém nagy** számítási és/vagy adattárolási igény jelentkezik.

Kik használhatják?

- Mindenki ingyen, akinek intézménye **tagi szerződéses** kapcsolatban áll az NIIF Intézettel.
- Egyetemi/főiskolai oktatók, hallgatók, kutatók, kutatóintézetek dolgozói, könyvtárak dolgozói, néhány kórház dolgozói.
- Cégek indirekt módon egyetemi partneren keresztül. **Csak kutatásra!**
- Kétlépcsős elektronikus regisztrációs folyamat:
 - Projekt regisztráció (csoport),
 - Felhasználói regisztráció (csoport felhasználói).
- Témaszám automatikusan érvényes az **összes** site-ra.

Regisztrációs űrlapkitöltő

<https://webform.niif.hu>

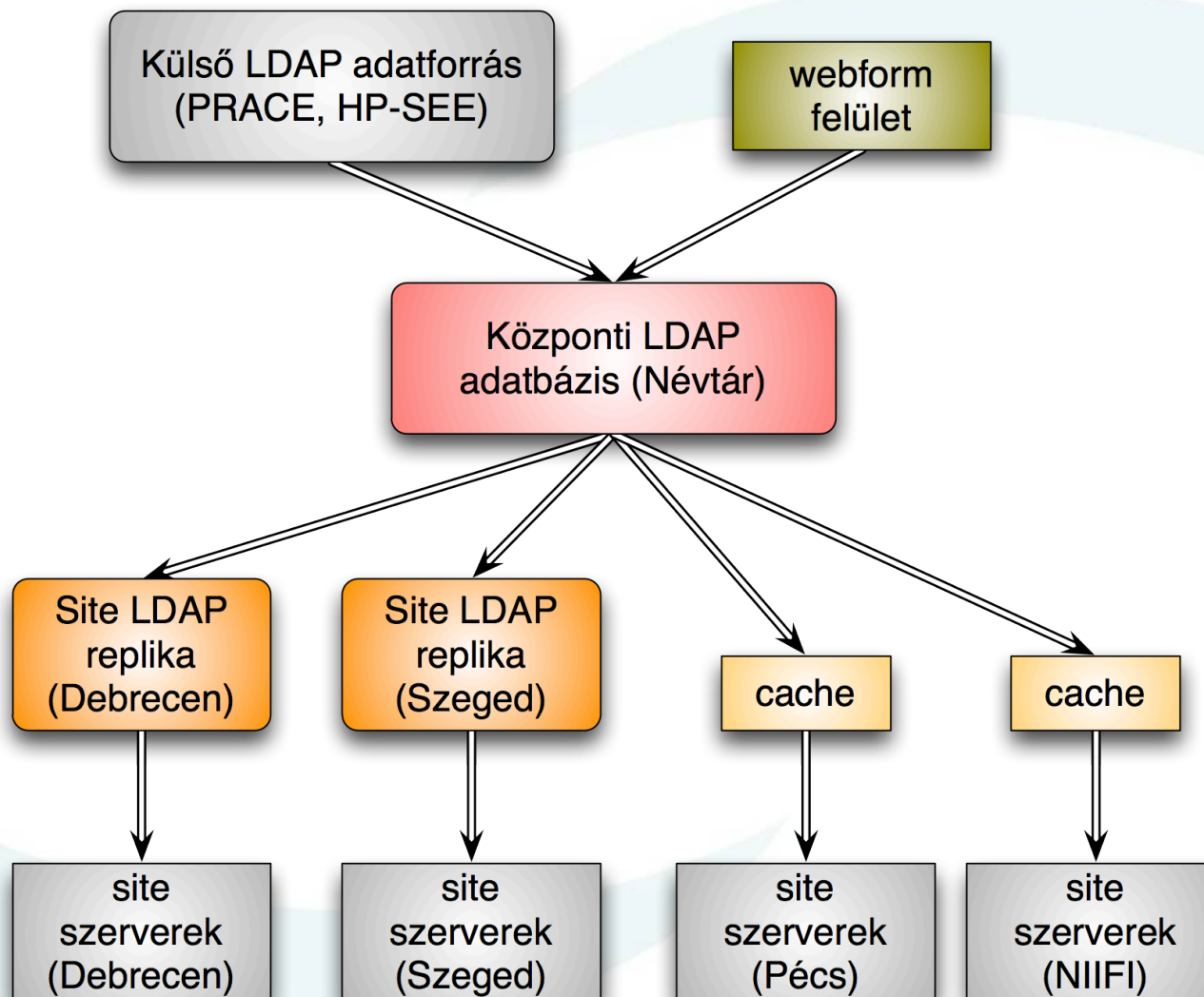


The screenshot shows the NIIF registration system interface. At the top, there is a blue header with the NIIF logo on the left and navigation links on the right: "Az NIIF Intézet szolgáltatásai", "Hogy működik?", and "Adatkezelési szabályzat". Below the header, the main title "Szolgáltatás igénylő és űrlapkitöltő rendszer" is displayed. The page content is titled "Címlap" (Homepage) and "Szupergép vagy grid projekt létrehozása, módosítása" (Creation or modification of supercomputer or grid project). A message states: "Az igénylés folyamatáról [itt talál leírást](#) Kérjük, töltsse ki a *-gal jelölt mezőket!" (For more information about the application process, see [here](#). Please fill out the fields marked with *!). Under the heading "Művelet:" (Operation:), there are two radio buttons: "Új projekt" (New project) and "Adatváltozás" (Data change). Below this, the "Projekt vezető" (Project leader) section contains three required text input fields: "Név: *" (Name), "Beosztás: *" (Position), and "Intézmény: *" (Institution).

Fontos felhasználási alapelvek

- Témaszám automatikusan az **összes site-ra**.
- SSH **kulcs alapú belépés**, nincsenek jelszavak!
- Frontend szerver.
- (Később grid köztesréteg alapú hozzáférés.)
- Elsősorban **köteget** használat.
- Helyi felhasználók priorizálása.
- Tudományos K+F+I célra használható, közvetlen jövedelemtermelésre **nem**.
- Külön file-rendszer az összes site-on, egyelőre kézzel kell adatot mozgatni.

Felhasználói azonosítás



Felhasználói támogatás

- Fontos célok:
 - Az országos HPC tudás aktív közösséggé történő összefogása.
 - Helyi tapasztalatok szintetizálása.
- Support kontakt:
 - helyi kontaktok,
 - szuper@niif.hu lista,
 - +36 1 4503095 bejelentő.
- További tervek:
 - Szervezeteken átívelő struktúra kialakítása.
 - Tudásbázis (wiki) kialakítása.

III. Belépés, kezelés, fordítás

Szoftverkörnyezet

- „Szoftver stack”:
 - operációs rendszer;
 - optimált matematikai könyvtárak;
 - párhuzamosításhoz használt könyvtárak;
 - fordítók, debuggerek;
 - felhasználói azonosítás (LDAP + replikák);
 - ütemezők;
 - grid köztesréteg;
 - alkalmazások (központi vs. user space);
 - mérő, monitorozó alkalmazások;
 - vizualizációs eszközök.

Belépés

- A belépés a site-okra SSH parancssori kliens segítségével történik:

```
ssh \  
user@login.debreceen.hpc.niif.hu  
ssh user@login.pecs.hpc.niif.hu
```

- Adatmásolás az SCP paranccsal lehetséges:

```
scp -r myfolder \  
user@login.debreceen.hpc.niif.hu:
```


Fordítók

- Elérhető a GNU, illetve a processzorgyártók fordítói, eszközei is:

`gcc`

`g++`

`g77`

`gfortran`

`icc`

`ifort`

`gmake`

- A standard, cc, f77, stb. fordító parancsok szimbolikus linkek.

Párhuzamos környezetek

- Ekkora erőforrás erejét **párhuzamos feldolgozással** lehet jól kihasználni:
 - tömb feladatok,
 - valódi párhuzamos feladatok.
- Támogatott párhuzamos könyvtárak:
 - MPI – üzenetküldés, el-, és megosztott memória, több implementáció telepítve;
 - PVM – üzenetküldés, el-, és megosztott memória, lassan kihalófélben van;
 - OpenMP – kernel folyamatok, megosztott memóriahasználat, általában már meglévő feladatok párhuzamosítására használják.

Fordítás

- Egyszerű fordítás:

```
gcc -o pi_gnu pi.c
```

```
icc -o pi_gnu pi.c
```

- Optimált fordítás:

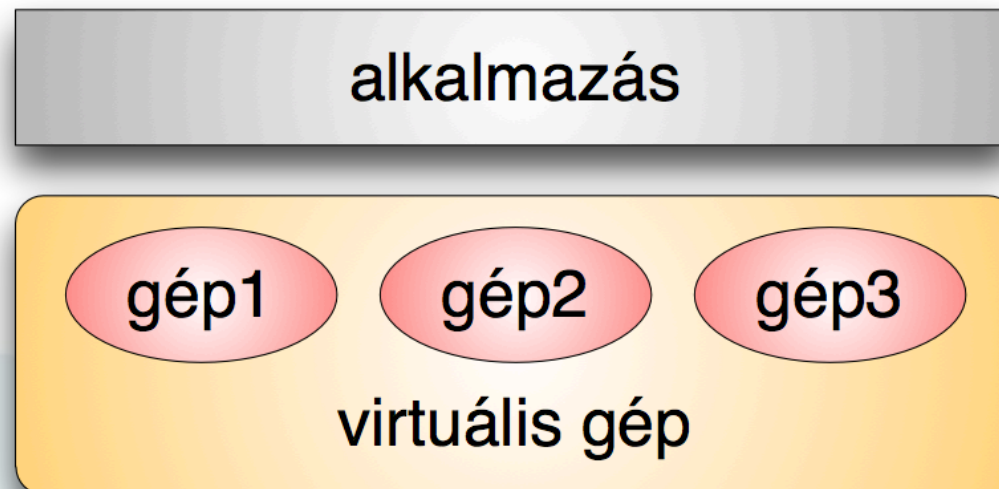
```
gcc -O5 -o pi_gnu pi.c
```

```
icc -O5 -o pi_gnu pi.c
```

- Programkód futtatásának hatékonysága függ a platformtól is.
- <https://www.niif.hu/node/314#3.7>

PVM architektúra

- Parallel Virtual Machine.
- Egy nagy virtuális gépet hoz létre hálózattal összekötött heterogén környezetben. Ezen futtatjuk az alkalmazásokat.
- Az adatcsere módja: üzenetváltás.



PVM fordítás

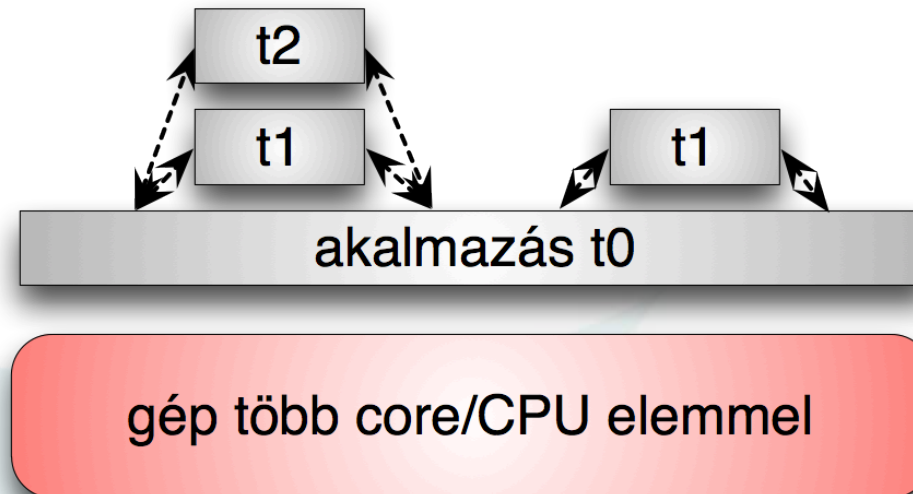
- PVM programok fordítása:

```
cc -o gexample gexample.c \  
-lpvm3 -dn \  
/usr/lib/pvm3/lib/LINUX64/  
libgpvm3.a
```

- Statikus vs. dinamikus linkelés összehasonlítása.
- Nem-standard PVM3 utak, szegedi, budapesti site-ok esetén (/usr/share).
- Idd, nm parancsok használata.

OpenMP architektúra

- Open Multi Processing Interface.
- Elosztott (shared) memóriás paradigma.
- Egyszerű programozói felület alkalmazások párhuzamosítására.
- Multiplatform nyelvi (preprocesszor), és runtime környezet kiterjesztés.



OpenMP fordítás

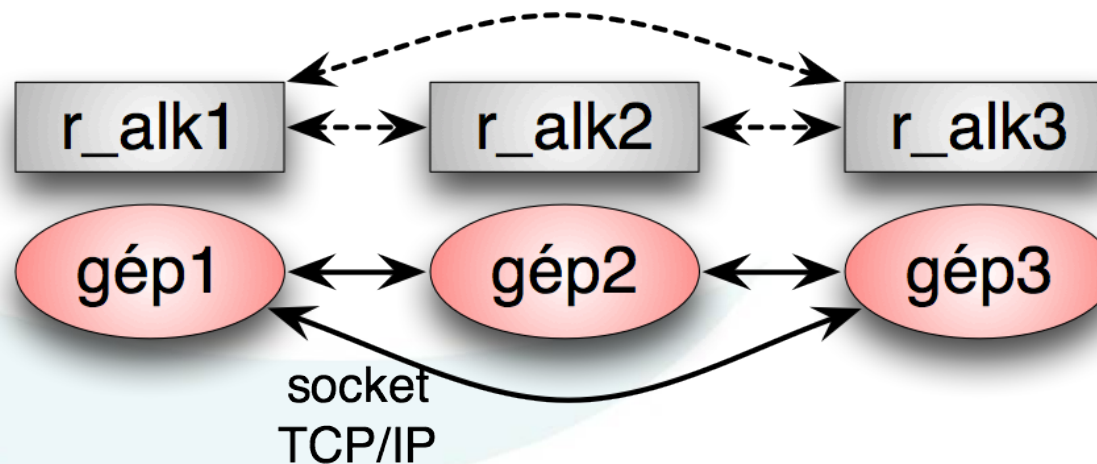
- OpenMP programok fordítása:

```
gcc -fopenmp -o omp_mm omp_mm.c  
icc -openmp -o omp_mm omp_mm.c
```

- Párhuzamos szálak beállítása a OMP_NUM_THREADS környezeti változó segítségével.
- Nincs hozzáfüzött könyvtár, a fordító végzi a párhuzamos kód kialakítását.

MPI architektúra

- Message Passing Interface.
- A számítási feladatot felbontja elemi részfeladatokra.
- Üzenetküldés.
- Standard felület, portabilitás, hatékonyság, kommunikációs topológiák.



MPI fordítás

- Több MPI változat, kiválasztani a megfelelőt:

```
mpi-selector --list
```

```
mpi-selector --set openmpi-1.3.2
```

```
module clear
```

- Kilépés-belépés.

```
mpicc -o connectivity \  
connectivity.c
```

- MPI parancsok = wrapper scriptek.

Make futtatása

- GNU make található az összes rendszeren.
- Default fordító környezeti változókkal (CC, F77, CXX, stb.) állíthatók be:

```
export CC=icc  
gmake -j 3
```

IV. Feladatok futtatása

Feladatok futtatása

- Minden site-on előttét szerverre léphetnek be a felhasználók.
- Direkt belépést tűzfalon tiltjuk.
- A feladatok köteget feldolgozása:
 - feladás,
 - sorbanállítás,
 - ütemezés,
 - futtatás,
 - eredmények kinyerése.
- Sun Grid Engine ütemező fogalmai:
 - queue, versus
 - host.

Feladatok futtatása

- SGE állapotának lekérdezése:

`qstat -f -u \*`

- Hostok lekérdezése:

`qhost`

- Párhuzamos környezetek:

`qconf -spl`

- Input állományok; output/error állományok keletkeznek.

Feladatok feladása

- Feladat (job) készítése:

```
#!/bin/sh  
#$ -N HOSTNAME_JOB  
/bin/hostname
```

- Feladat feladása:

```
qsub script.sh
```

- Feladat törlése:

```
qdel job_id
```

Feladatok feladása

- Feladat módosítása:

```
qalter -pe pvm 12 job_id
```

- Feladat hibaállapotának törlése:

```
qmod -c job_id
```

- Feladat futásának felfüggesztése:

```
qmod -s job_id
```

```
qmod -us job_id
```

Tömb feladatok feladása

- Tömb feladat futtató script:

```
#!/bin/sh
#$ -N PI_ARRAY_TEST
./pi_gcc \
`expr $SGE_TASK_ID \* 100000`
```

- Feladat feladása:

```
qsub -t 1-7 script.sh
```


Párhuzamos feladatok feladása

- Parancs:

```
qsub -pe mpi 21 script.sh
```

- A párhuzamos környezet lehet: pvm, mpi, openmp.
- Trükkök script beljében (mpirun, OMP_NUM_THREADS beállítása).
- \$NSLOTS változó a script belsejében.
- Parancssori attribútumok a scriptben: \$\$ makrók.

V. Záró gondolatok

Amiről nem esett eddig szó...

- Checkpointing/restart támogatás:
 - felhasználói,
 - könyvtár-szintű,
 - kernel-szintű.
- Vizualizáció.
- Grid köztesrétegeken keresztül történő futtatás.
- AAI integráció.
- Installált csomagok listája:
`rpm -qa`
`rpm -qi csomagnév`
`rpm -qf filenév`

Tervezett bővítések

- Pénzes szoftverek:
 - Gaussian,
 - Matlab,
 - Maple.
- Grafikus kártyákat tartalmazó csomópontok. (Újabb csomópontok.)

Összefoglalás, trendek

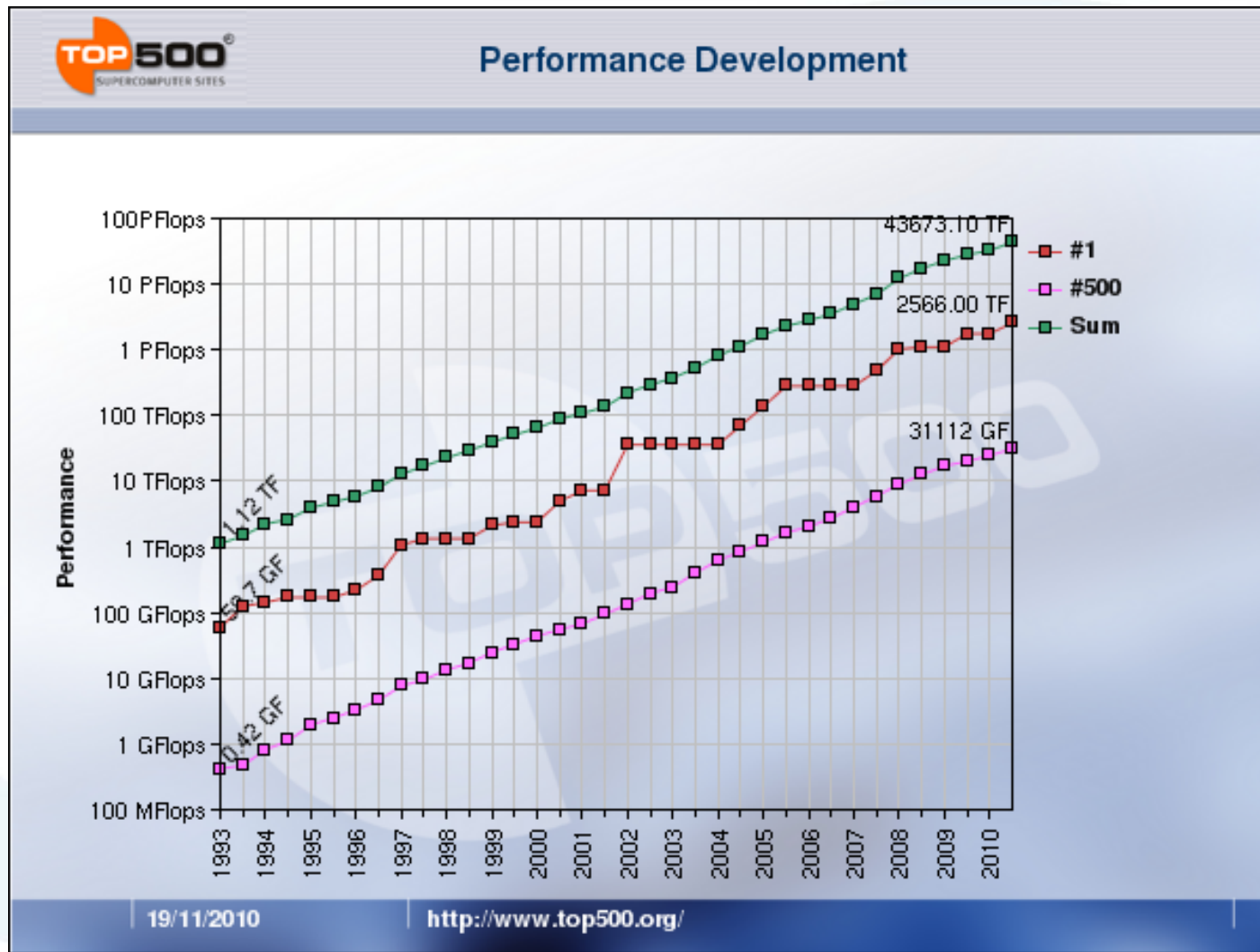
- A rendszer összjellemezői:
 - Kapacitás: 47.85 Tflops (100 x);
 - Magok száma: 4780 darab (20 x);
 - Memóriaméret: 8 Tbyte (20 x);
 - Diszk kapacitás: nettó 1033 Tbyte (50 x);
 - Összekapcsolás: NIIFI HBONE+ 10GB/s (10x);
- Összköltség: bruttó 1.018 mrdFt.
- A legutóbbi Top500 **343. helyére** kvalifikálhatnánk magunkat vele.
- Szuperszámítástechnika nemzetközi és hazai helyzete, nemzetközi integráció (PRACE, EGI, HP-SEE)...

Összefoglalás, trendek

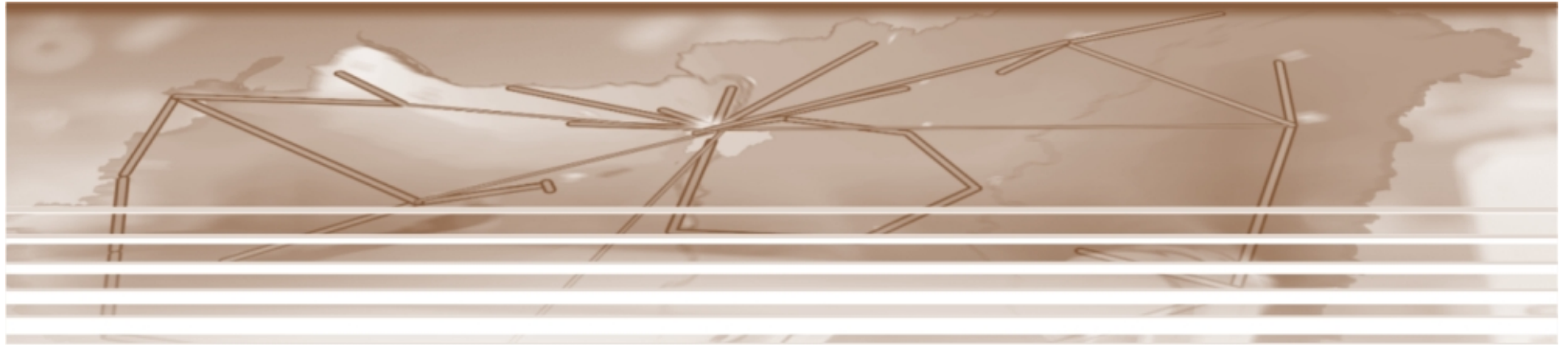
- Trendek:

- A csúcs kiszolgáló rendszerek kapacitása folyamatosan növekszik.
- Ennél jobban csak a felhasználói igények.
- Mainframe és PC technológia összesimulása.
- Vendor-specifikus processzorok kikopása.
- Fürtözött rendszerek jelentős előretörése.
- GPU-k jelentős előretörése.
- Virtualizáció erősödése: HPCaaS (Intel 56xx-as széria).
- Új datacenter összeköttetések: 40GE, 100GE.
- Linux dominancia, feltörekvő MS HPC környezet.

Összefoglalás, trendek



Köszönöm a figyelmet!



Stefán Péter
<szuper@niif.hu>