

Az új projekt és GPU igénylési lehetőségek áttekintése a HUN-REN Cloudon

Kacsuk Péter
HUN-REN SZTAKI

2025. 01. 30.



Miért kellett új projekt és GPU igénylési lehetőségeket bevezetni?

- Az **AI előretörésével meredeken nőtt az igény** a GPU-t használó projektek indítására, így a kiosztható GPU kapacitás elfogyott
 - főleg a hosszabb, így gyakran hektikus/részleges kihasználtságú projektek miatt
- Egyre nagyobb igény mutatkozik még **újabb és kényelmesebb AI szolgáltatásokra**, ld. GenAI4Science
- Köszönhetően (többek között) az AI4Science programnak növekednek az AI kompetenciák, egyre **haladóbb és nagyobb felhasználási esetek** jelentek meg, ami más (kötegetelt, job-alapú) megközelítést igényelnek
- A felhő **következő bővítésének ütemezése elhúzódik: 2026**

A GPU használati módok kibővítése új lehetőségekkel

1. IaaS mód (jelen előadás és Farkas Attila első előadása)
 - 1 éves igényű projekt
 - **Szolgáltatást nyújtó projekt**
 - **Ernyőprojekt**
 - 3 hónapos igényű projekt
 - Előkészítő projekt
 - Haladó projekt
2. **PaaS mód SLURM job manager alapon (Ld. Emődi Márk előadását)**
 - Interaktív üzemmód
 - Batch üzemmód
3. **GenAI4Science platform használata (Farkas Attila 2. előadása)**

Új GPU elosztási rendszer

Az új lehetőségek mellett új GPU elosztási rendszer kialakítása is szükségessé vált

Olyan új rendszer bevezetése volt a cél, amely

- Elegendő erőforrás esetén úgy működik, mint a jelenlegi rendszer
- Ha viszont nincs elegendő erőforrás, akkor automatikusan átmegy egy a versenyhelyzetet fair módon lekezelő mechanizmusba

A versenyhelyzet kezelése 2 szempontból is fair:

1. Nem kell adott esetben egy évet várni a projekt indítására, legkésőbb 3 havonta elbírálásra kerül minden projektigény
2. A versengő projektigények egységes szempontok (tudományos teljesítmény, impakt, erőforrásigény, stb.) alapján kerülnek elbírálásra



Az új GPU elosztási rendszer fogalmai



1 éves projektek

Az 1 éves **szolgáltatást nyújtó projekt** olyan projekt melyben folyamatos szolgáltatást nyújt a projekt külső felhasználók számára (pl. arcfelismerő szolgáltatás).

- A HUN-REN Cloud „best effort” szolgáltatást nyújt SLA nélkül, tehát a cloudon futó szolgáltatást nyújtó projektek is csak ilyen szolgáltatást tudnak nyújtani

Az 1 éves **ernyőprojekt** egy olyan projekt, amelyben nem egy, hanem több projekt/kutatócsoport osztozik az igényelt erőforrásokon és az erőforrások kiosztását, egymás közötti ütemezését saját maguk szervezik meg. Pl. egy egyetemi kar projektjei együttesen használják a számukra allokált GPU erőforrásokat

3 hónapos előkészítő projekt

A 3 hónapos **előkészítő projekt** egy olyan projekt, ami most kér először GPU erőforrásokat a felhőn és a résztvevő kutatóknak még nincs tapasztalatuk a GPU használatával a felhőn.

- Az ilyen projektek tipikusan azon dolgoznak, hogy kialakítsák azt a szoftver környezetet, amiben a GPU-kat használni fogják. Ez általában egy időigényes feladat és nem jár együtt intenzív GPU használattal. Ezért az előkészítő projektek csak **a legkisebb GPU flavort igényelhetik 1 GPU-val**.
- A 3 hónap kitöltése nem kötelező. Ha a projekt előbb elkészül a szükséges környezet kialakításával, akkor az előkészítő projektet lezárhatja és egy új haladó projektet igényelhet (ld. haladó projekt igénylési űrlap).

3 hónapos haladó projekt

A 3 hónapos haladó projekt lehet:

egy sikeres előkészítő projekt	egy teljesen új projekt , ahol a projekt valamelyik közreműködője már rendelkezik GPU tapasztalattal a HUN-REN Cloudon és ezért azonnal tudják használni a cloud GPU szolgáltatásait	egy haladó projekt
▪ azonnali folytatása, ▪ vagy megszakítási idő utáni folytatása	▪ ilyenkor az új projekt kérő űrlapon meg kell indokolni, hogy miért nincs szükség előkészítő projektre és röviden le kell írni a korábbi cloud GPU szolgáltatás használati tapasztalatokat	▪ azonnali folytatása, ▪ vagy megszakítási idő utáni folytatása

A 3 hónap jelentése

- A 3 hónap azt jelenti, hogy azok a projektek, amelyek nem igényelhetnek egyszerre 1 éves időtartamot a felhőn (mivel nem nyújtanak szolgáltatást és nem ernyőprojektek, hanem egyedi projektek) egyszerre csak 3 hónap időtartamra kérhetik a felhő használatot.
- Ugyanakkor, ha úgy látják, hogy a 3 hónap nem lesz elég a projekt befejezéséhez, akkor kérhetik a projekt hosszabbítását, de ez az igény is csak 3 hónapra szólhat.
- Viszont a 3 hónapos hosszabbítási igény akárhányszor beadható.

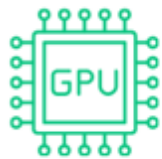
Fogalmak a projekt igények eldöntésére

A GPU allokálási rendszer kétféle állapotban lehet:

- **Alapállapot** (van elég GPU kapacitás az aktuális projekt igény kielégítésére)
- **Versenyállapot** (nincs elég GPU kapacitás az aktuális projekt igény kielégítésére)

Megkülönböztetünk háromféle időszakot:

- **3 hónapos Projekt időszakok** (a 3 hónapos projektek végrehajtási időszaka)
 - (jan. 1. – márc. 31.), (ápr. 1. – jún. 30.), (júl. 1. – szept. 30.), (okt. 1. – dec. 31.)
- **Elbírálási időszakok** (a várakozó projekt igények elbírálási időszaka)
 - márc. 1-14, jún. 1-14, szept. 1-14., dec. 1-10.
- **Felkészülési/zárási időszakok** (nyerteseknek felkészülés a következő projekt időszakra, többieknek projekt zárás)
 - márc. 15-31, jún. 15-30, szept. 15-30., dec. 11-23.



Igény felmerülése

Igénylő űrlap
beküldése a portálonGPU igénylőlap
csatolásával

Pályázat benyújtása



PROJEKT FUTTATÁSI IDŐSZAK



Szabad kapacitások áttekintése

Pályázó értesítése a pozitív
elbírálásrólIgény felfüggesztése az
elbírálási időszak kezdetéigHUN-REN Cloud
projekt futtatása(Előző elbírálási időszakban jóváhagyott
HUN-REN Cloud projekt futtatása)HUN-REN Cloud
projekt futtatásaÚjrapályázás
megfontolása/ egyéb
megoldás kereséseBeadási
határidőElbírálás
végeÚj projektek
indulása

- Tudományos érték?
- Szabad erőforrások?
- Hosszabbítás esetén
eddig elért
eredmények?
- Az új- és hosszabbítási
igények azonos súllyal
esnek latba

Igény elbírálása

Pályázó értesítése a pozitív
elbírálásrólPályázó értesítése az
elutasításrólAktuális időszakban futó
projektek befejezése/zárása,
akik nem pályáztak újra
hosszabbítás céljából vagy
pályáztak, de elutasításra
kerültek

ZÁRÁSI IDŐSZAK

PROJEKT FUTTATÁSI IDŐSZAK

További lehetőségek

Ha egy projekt igény nem nyert, vagy nem akar várni a következő Elbírálási időszakig

1. Ha úgyis szakaszos mintát követ a tervezett GPU használat, akkor érdemes megfontolni a **SLURM job manageren** keresztül elérni a GPU erőforrásokat (itt akár egy hétre meg lehet kapni a GPU erőforrásokat)
2. Ha generatív AI feladat megoldására kell a GPU, akkor érdemes megfontolni a **GenAI4Science platform használatát**



A projekt pályázatok elbírálása és a GPU igénylőlapok

- Az elbírálás érdekében minden új és hosszabbítási projekt igényhez csatolni kell egy kitöltött GPU igénylőlapot a felhő weblapján (ld. Farkas Attila első előadását).
- A projektek elbírálásában az itt megadott információkra támaszkodunk.
- Az új és hosszabbítási igények azonos súllyal esnek latba (egyik sem preferált)

Ökölszabályok az új rendszerben

- Ha van futó projekted és akarod folytatni, akkor ne feledd el a projekt hosszabbítási igényt beadni az aktuális Elbírálási időszak kezdetéig (ha ezt elmulasztod, mindenképp zárnod kell a projektet a Projekt időszak végére)
- A projektedre akárhányszor lehet hosszabbítást kérni
- Új projekt igény bármikor beadható!
- Ha van elég erőforrás az új projekt azonnal indítható (úgy mint régen)
- Ha nincs elég erőforrás, akkor meg kell várni a következő Elbírálási időszakot
- Mindenre kiterjedő részletes magyarázat található a weblapon elérhető **GPU Erőforrás-gazdálkodási Szabályzat**ban:

<https://science-cloud.hu/gpu-eroforras-gazdalkodasi-szabalyzat>



Segítség a GPU igénylőlapok kitöltéséhez és az új GPU Erőforrás-gazdálkodási Szabályzat megértésében

<https://science-cloud.hu/support>

 **HUN-REN Cloud** [Rólunk](#) [Aktualitások](#) [Projektek](#) [Dokumentáció](#) [GenAI4Science](#)

[← Címlap](#) | [Támogatás kérése](#)

Támogatás kérése

Vezetéknév *

Keresztnév *



Köszönöm a figyelmet!

Bármilyen kérdés?

