Felhő-erőforrás menedzsment

Toka László

SINCE 1992

HSN Lab

- Bemutatkozás
- Célok
- Megbízhatóság
- Skálázás
- Csomagolás

Bemutatkozás

- PhD Telecom ParisTech 2011
- Ericsson Research 2011-2014
- MTA-BME Informatikai Rendszerek Kut.csop. 2014-2022
- ELKH-BME Felhőalkalmazások Kut.csop. 2022-
- Felhő, AI, infokomm
- BME VIK TMIT docens
- HSNLab vice-head
- 2 Bolyai, MTA DSc bírálat alatt

HSNLab SINCE 1992

MŰEGYETEM 1782

HUN-REN Magyar Kutatási Hálózat

MAGYAR TUDOMÁNYOS AKADÉMIA 1825

HSNLab SINCE 1992

Célok

- „Olyan rendszerek és módszerek tervezése, amelyek a felhőalapú számítástechnika világát még megbízhatóbbá, megfizethetőbbé és könnyebben elérhetővé teszik az ügyfelek számára világszerte”
- Felhőkezelési technikákat javaslok a magas szolgáltatásminőség (QoS) érdekében
- A legfontosabb elvárt QoS metrikák
  - Megbízhatóság → redundancia
  - Rendelkezésre állás → skálázás
  - Gyorsaság → csomagolás
- Költségminimalizálás

DSc Téziscsoport 2

- Javasoltam egy rendszert és benne működő módszereket, amelyek **magas megbízhatóságot** és **alacsony késleltetést** biztosítanak **gazdaságos módon** nagyléptékű **peremfelhőkben**. Egy online és egy offline erőforrásütemező, és egy peremfelhő csomópont-szegmentáló eljárás kerül bemutatásra, amelyek együttesen nagy léptékben képesek kezelni a szolgáltatástelepítési kéréseket egy földrajzilag kiterjedt felhő infrastruktúrában.
- Javasoltam továbbá egy gépi tanulás-alapú automatikus **skálázási** módszert annak érdekében, hogy kezelni tudjam az online alkalmazások használati intenzitásának időbeli változékonyságát.
- Elemeztem a felhőalapú alkalmazások különböző **telepítési lehetőségeinek** hatását a válaszidőkben jelentkező késleltetésre és a memóriafogyasztásra, és javasoltam egy költséghatékony módszert mikroszolgáltatások futtatására.

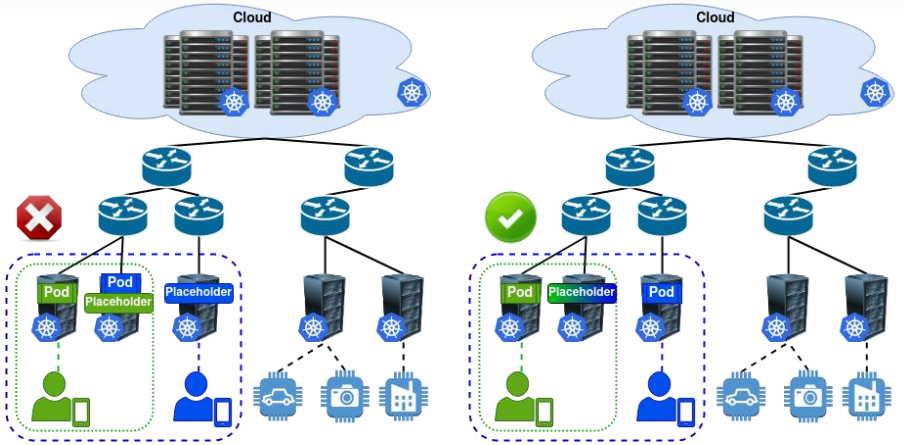
Megbízható peremfelhő

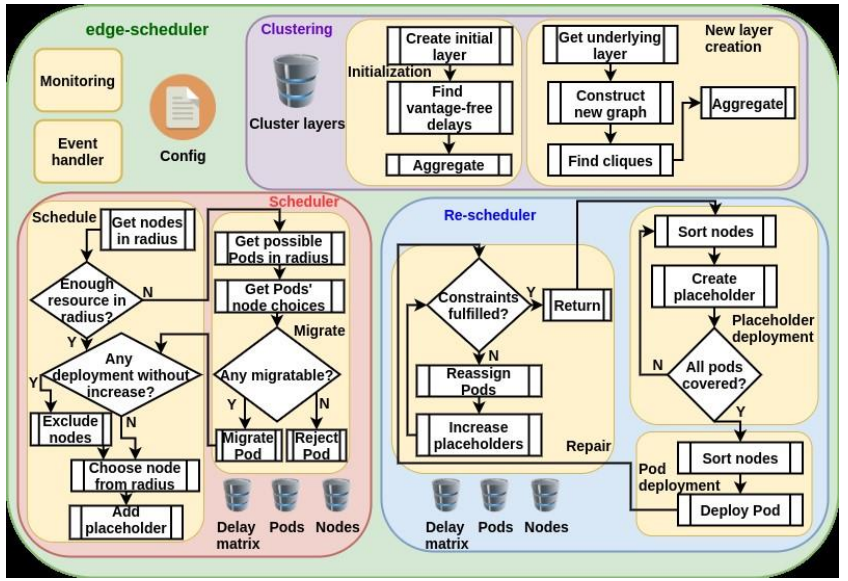
- peremfelhő ütemezés: a szerverek javítása/cseréje lassú, a redundancia költséges és egy helyszínen hiábavaló lehet

- biztonsági mentési erőforrások bevezetése – „helyőrzők”

- helyőrzők kapacitása 1-csomóponti meghibásodásokra (azaz bármely csomóponton az összes Pod egyszerre meghibásodhat)

- a helyőrzők kiszámítására és a Pod-ok elhelyezésére javasolt algoritmusok gyorsnak bizonyultak.





Laszlo Toka. Ultra-reliable and low-latency computing in the edge with Kubernetes. Journal of Grid Computing, 19(3), 2021.

Költségkímélő skálázás

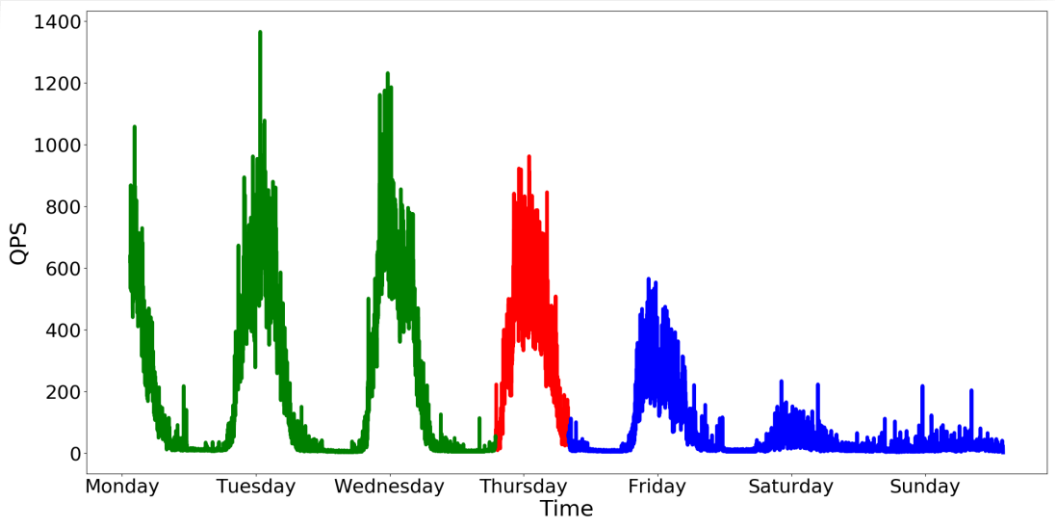
- túlskálázás: költség, alulskálázás: QoS romlás

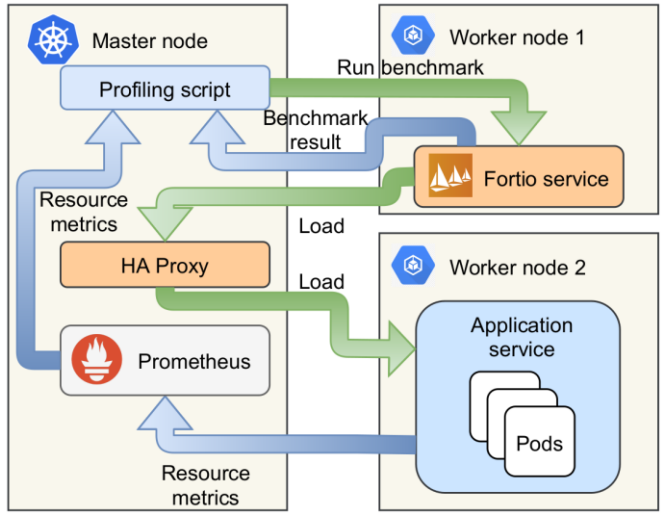
- szimulációs összehasonlítások céljára: veszteségmentes modell a kérések érkezési folyamatát Markov-modulált Poisson folyamatként (MMPP) írja le, veszteséges modell egy összetettebb, diszkrét idejű sorbanállási modell

- a bemeneti kérések általában változó dinamikát mutatnak a nap folyamán, ezért a különböző MI módszerek epizodikusan jól vagy rosszul teljesítettek

- gépi tanulási modellek kombinálása az alkalmazáshasználati előrejelzésekhez: ensemble modell

- gépi tanulás alapú ensemble modellre épülő prediktív automatikus skálázási módszer jelentős költségmegtakarítást ér el (akár 50%) az alapértelmezett benchmarkhoz képest, főként az elveszett kérések jelentős csökkenésének köszönhetően





Laszlo Toka et al. Machine learning-based scaling management for Kubernetes edge clusters. IEEE Transactions on Network and Service Management, 18(1):958–972, 2021.

HSN Lab
SINCE 1992

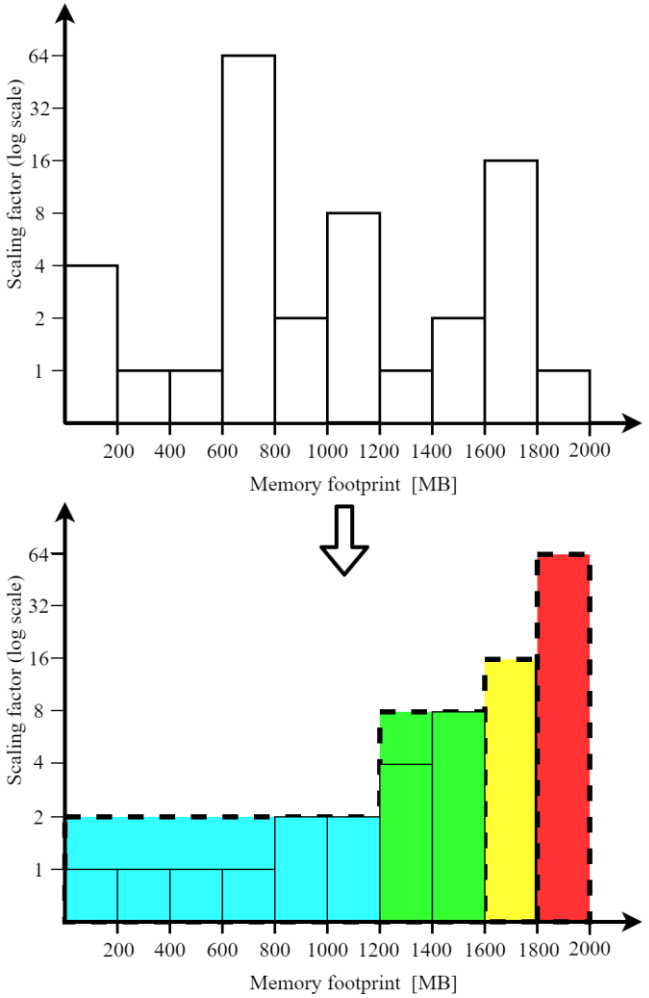
Csomagolás: költség vs késleltetés a skálázási egységek miatt

- A skálázás költségét nagymértékben meghatározza az alkalmazás skálázási egységekre való szervezése.

- Egyrészt az alkalmazás-összetevők közös skálázási egységben való elhelyezése alacsonyabb működési késleltetéseket eredményez, így jobb QoS biztosítható az alkalmazás felhasználói számára.

- Másrészt a kisebb modularitás felesleges erőforrás-felhasználást eredményez a nagyléptékű működési időszakok során

- analitikai modell az erőforrás-többlet és a késleltetési többlet közötti döntési helyzet leírására.



Laszlo Toka. The shape of your cloud: How to design and run polylithic cloud applications. IEEE Access, 10:97971–97982, 2022.

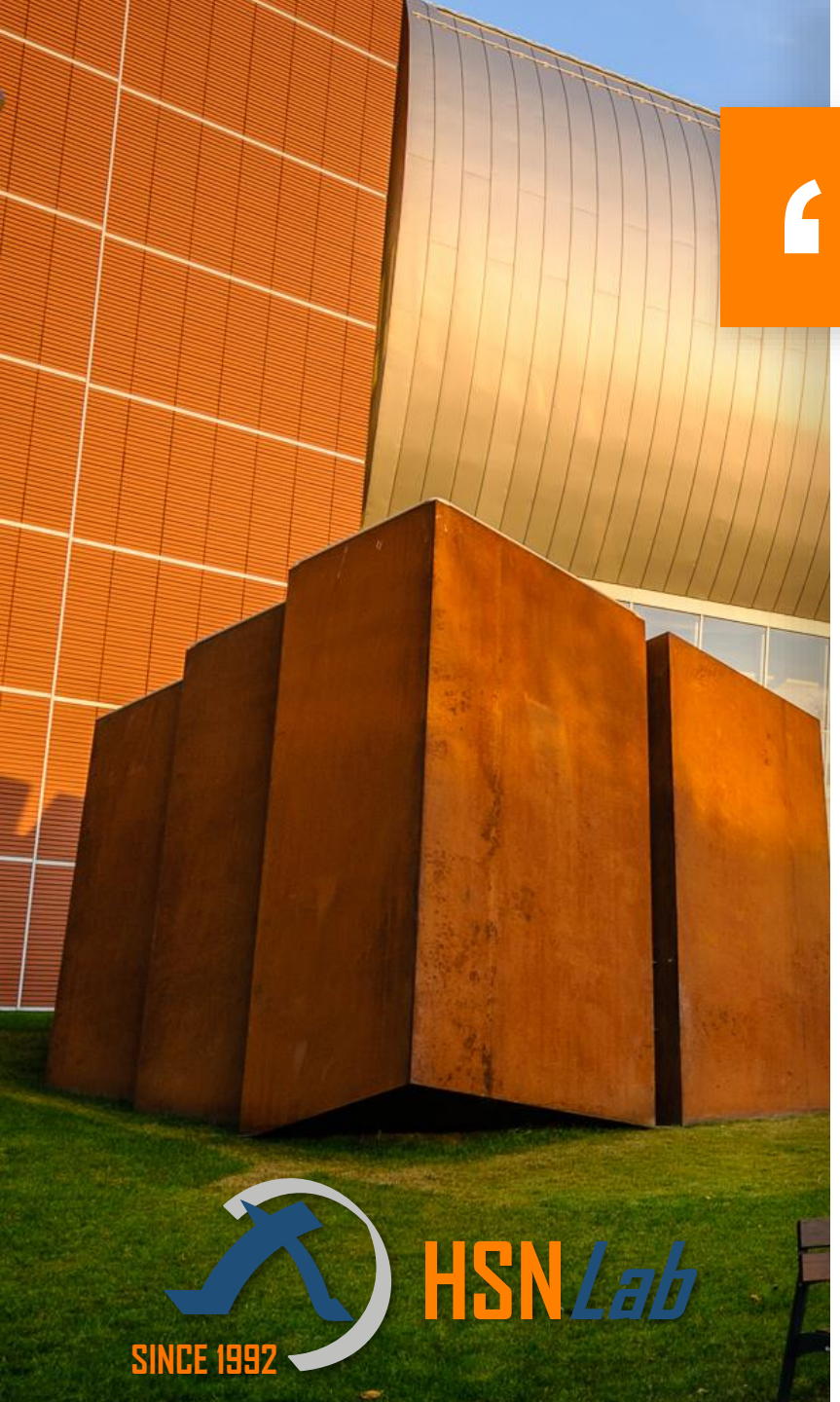
HSN Lab
SINCE 1992

Kapcsolódó konferencia cikkek

[C3] Laszlo Toka, David Haja, Attila Korosi, and Balazs Sonkoly. Resource provisioning for highly reliable and ultra-responsive edge applications. In IEEE International Conference on Cloud Networking (CLOUDNET), 2019.

[C4] David Haja, Mark Szalay, Balazs Sonkoly, Gergely Pongracz, and Laszlo Toka. Sharpening Kubernetes for the Edge. In ACM SIGCOMM Conference Posters and Demos, 2019.

[C5] Laszlo Toka, Gergely Dobreff, Balazs Fodor, and Balazs Sonkoly. Adaptive AI-based auto-scaling for Kubernetes. In IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID), 2020.



Köszönöm a figyelmet!

toka.laszlo@vik.bme.hu