



BUDAPESTI MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM
VILLAMOSMÉRNÖKI ÉS INFORMATIKAI KAR
MÉRÉSTECHNIKA ÉS INFORMÁCIÓS RENDSZEREK TANSZÉK

Parciális rekonfiguráció

Heterogán számítási

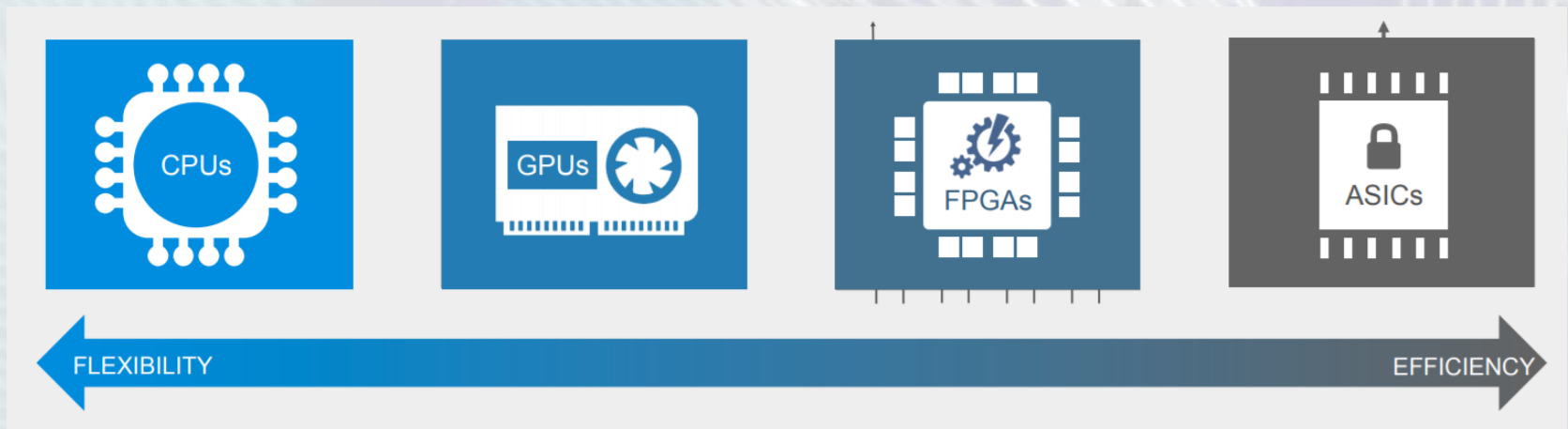
rendszerek

VIMIMA15

Fehér Béla
BME MIT

Felhőszolgáltatások igényei

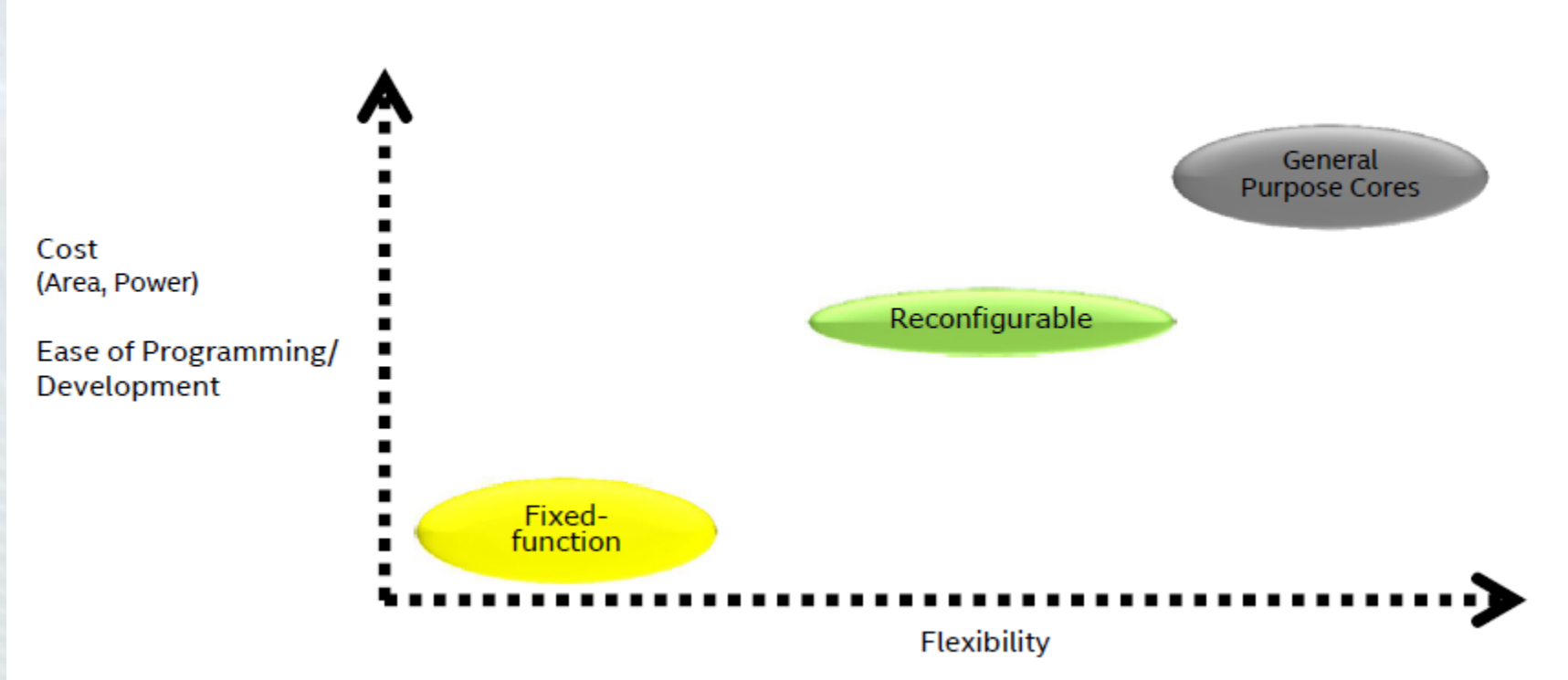
- Nagyságrendi növekedés az adatforgalomban
- Magasabb követelmények a válaszidők, késleltetés (latency) csökkentésére
- Nagyobb feldolgozási sebességigény
- **Alternatívák: CPU-GPU-FPGA-ASIC**



Alternatívák és opciók

- **Működési hatékonyság: Művelet/W, Művelet/\$**
- **Programozási hatékonyság: Erőforrás, Költség**

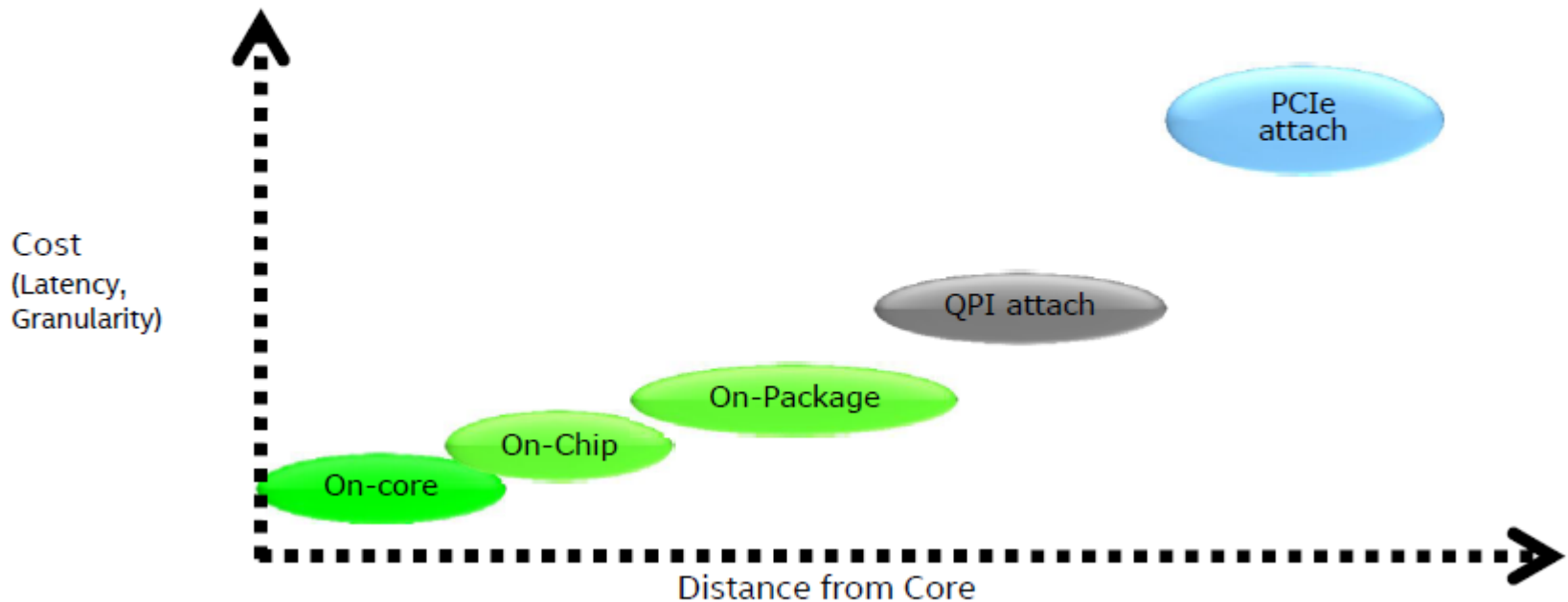
Accelerator Architecture



Alternatívák és opciók

- Rendszerbeillesztés helye, módja,
- Kapcsolat a Host és a gyorsító között

Accelerator Attach

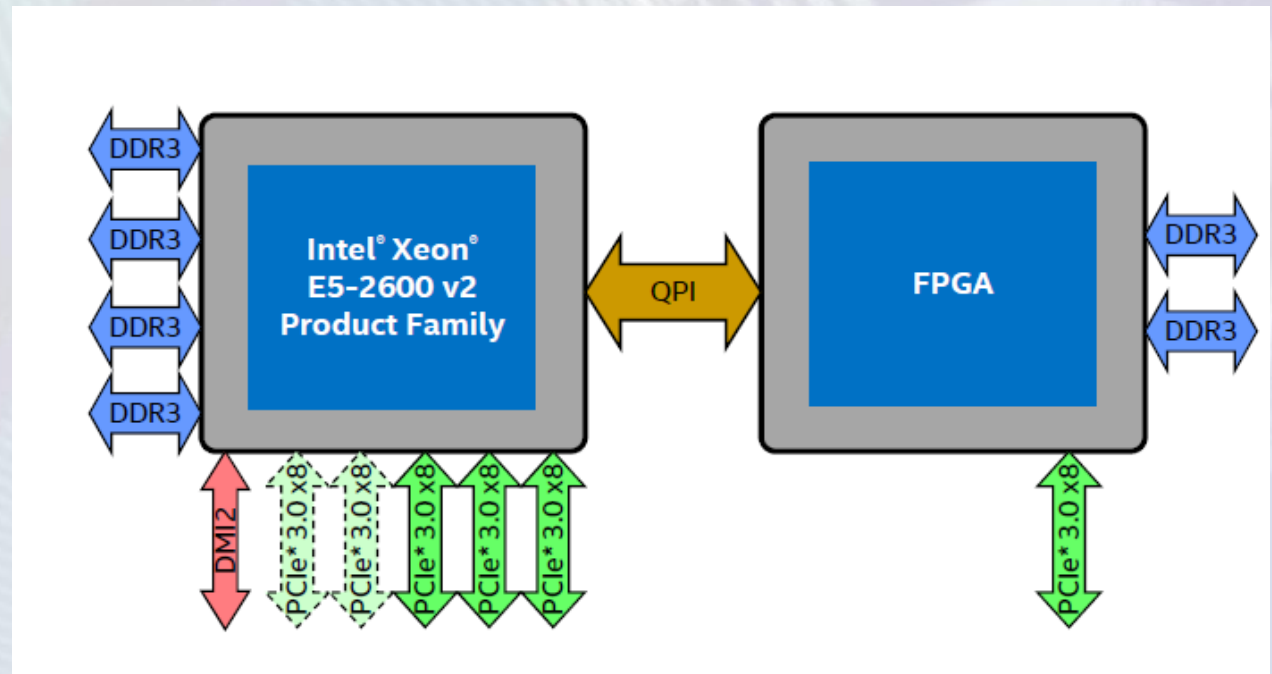


Alternatívák és opciók

- **Programozási és használati modell**
- **Adathozzáférési modell**
 - Közvetlen IO csatornán keresztül
 - Megosztott virtuális memóriában
 - Nem kell adat buffereket másolgatni...)
- **Együttműködési modell**
 - Off-load: független működés
 - Hibrid: az algoritmus részei megosztva működnek a host és a gyorsító erőforrásain

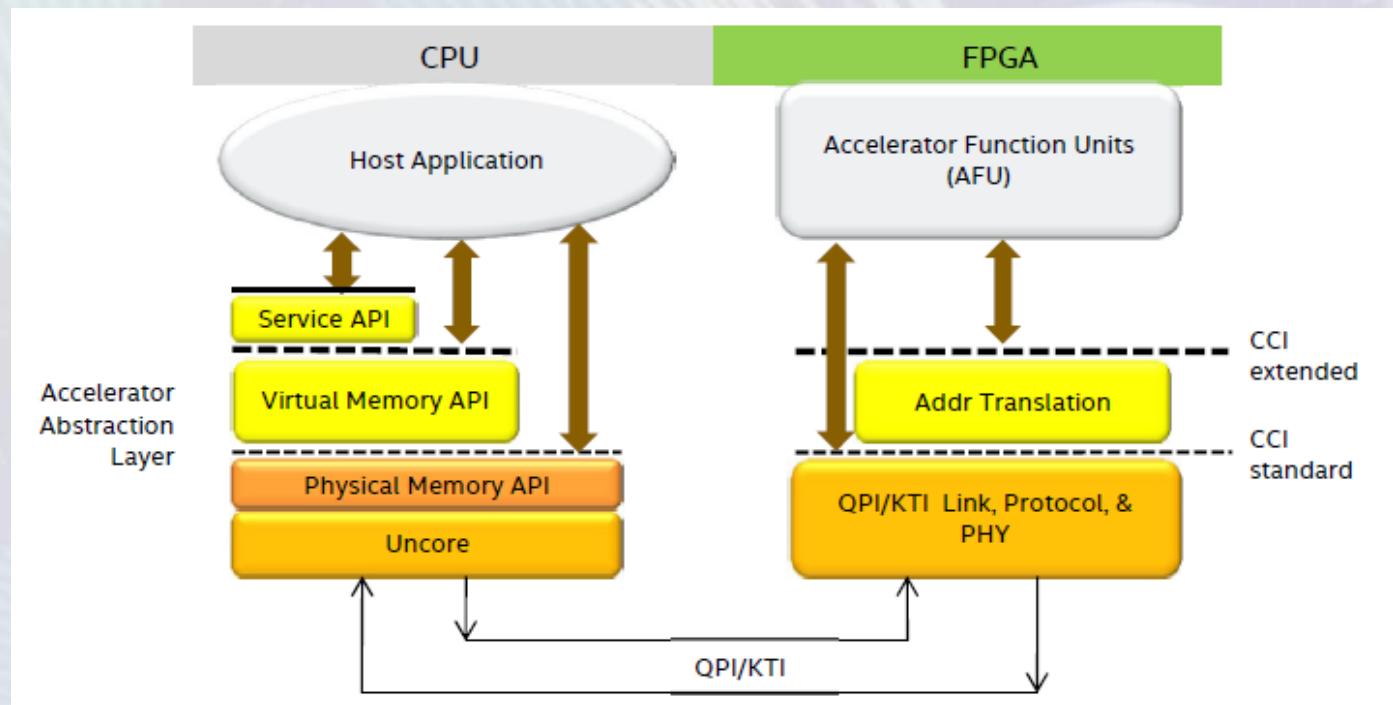
Intel megoldás

- **Xeon + FPGA QPI interfésszel**
 - Külső kapcsolat QPI 6,4GT/s sebességgel
 - Lokális memória, bufferelt adatkezelés
 - AAL Acceleration Abstraction Layer



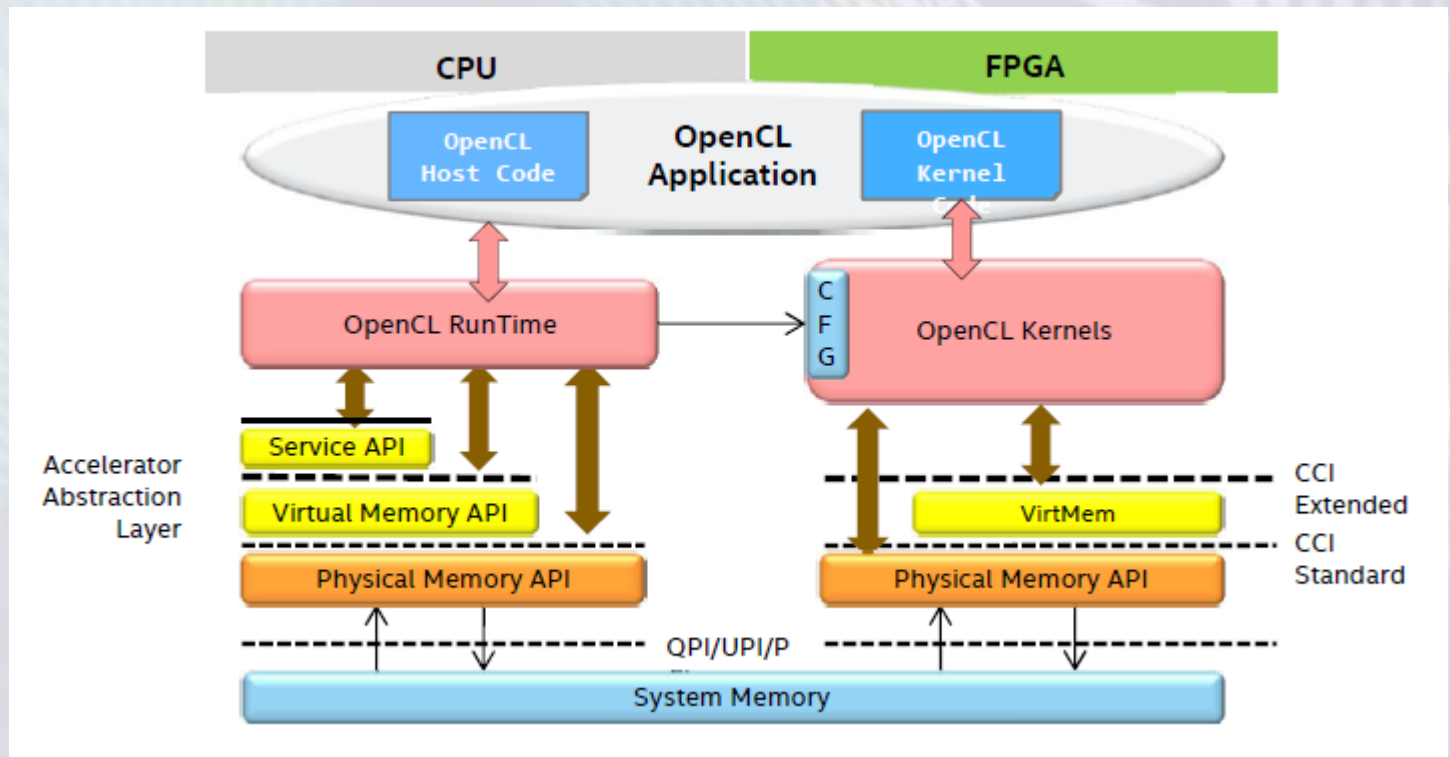
Intel megoldás

- **Programozási interfész**
 - AFU Accelerator Functional Units
 - CCI Core Cache Interface (másoknál CCIX)
 - Egységes memória kezelés a teljes rendszerben



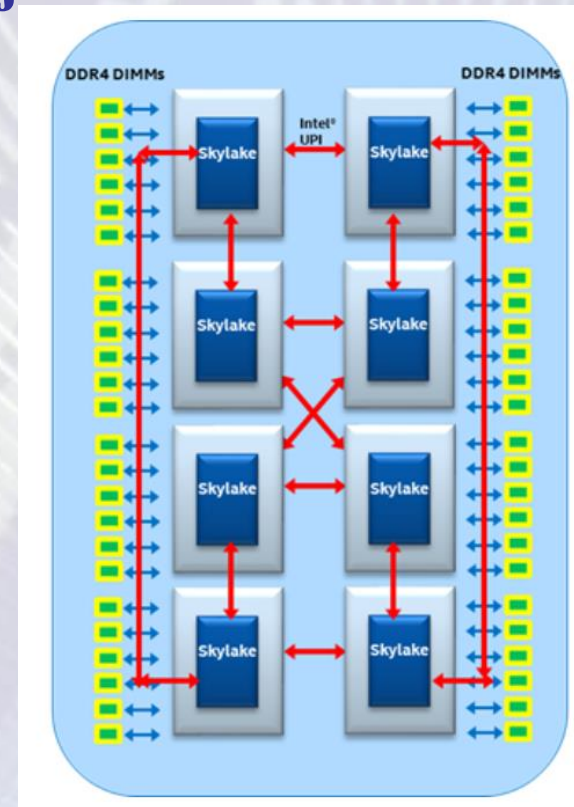
Intel megoldás

- Programozási interfész OpenCL esetén
 - Egységes alkalmazói kód
 - Hordozható CPU+FPGA generációk között



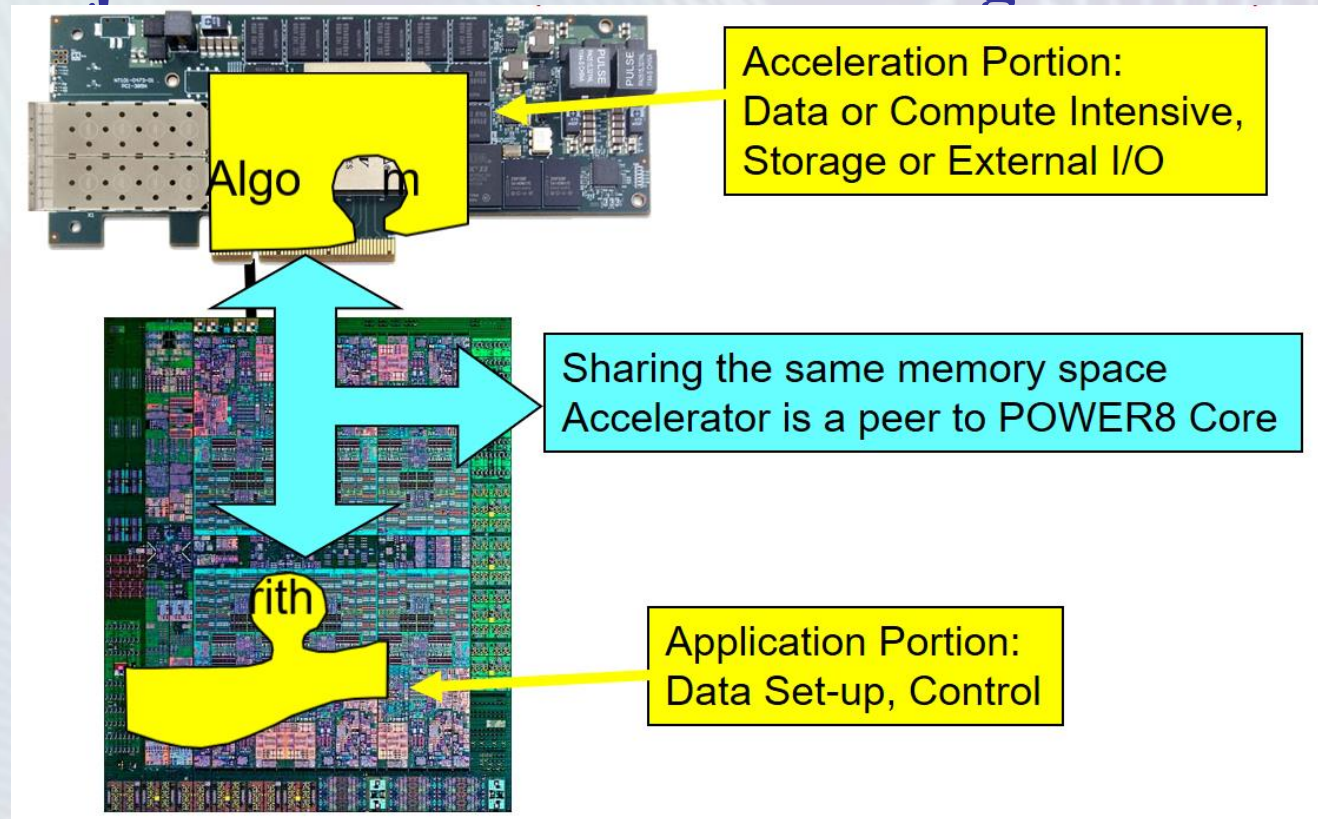
Intel UPI

- Ujabb generációs CPU-k Skylake és utána
- UPI Ultra Path Interconnect – Teljes 2D hálózat
- Cache-koherens kapcsolat
- Elsősorban a core-ok között
- De külső egységek is felfűzhetők
 - Gyorsítók
 - Speciális tároló egységek



IBM CAPI

- CAPI Coherent Accelerator Processor Interface
- Szabványos PCIe fizikai kapcsolat, a POWER8/9 CPU biztosítja a rendszerbe illeszthetőséget



IBM CAPI

- **A cél a gyorsító használatát előkészítő felesleges műveletek (pl. memória adatmásolás) kiiktatása**
- **Koherens memória elérés nélkül az eszközkészítő felelős a memória leképezéséért, változók, bemeneti adatok mozgatásáért**
- **Az adatok egyidőben több példányban léteznek**
 - Eredeti (OS) adatok a rendszermemóriában
 - Átmeneti adatok az eszközkészítő puferrében
 - Munka adatok a gyorsító lokális memóriájában
- **Sok felesleges adminisztráció a driver-ben**

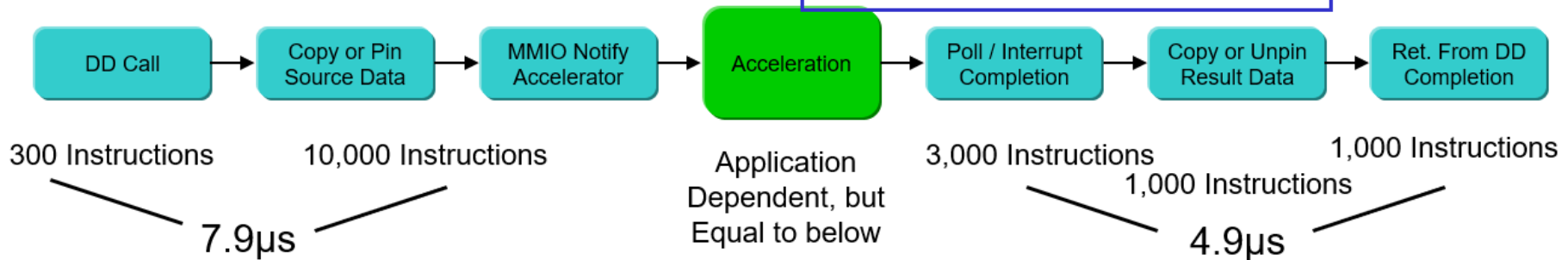
IBM CAPI

- A CAPI megoldásban definiáltak egy PSL POWER Service Layer-t
- Ez a HW modul az FPGA-ban kerül kialakításra, közvetlenül a POWER8/9 CPU végrehajtó egységével/memória vezérlőjével áll kapcsolatban
- Nincs felesleges adatmozgatás, a gyorsító a változókat, bemeneti adatokat közvetlenül a rendszer memóriából éri el, és az eredményt oda írja vissza
- Az eszközkezelő a működés során nincs használava

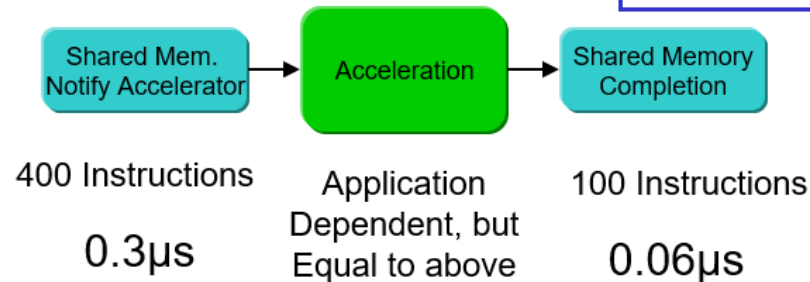
IBM CAPI

- Az adatelőkészítés folyamata

Typical PCIe Model Flow: Total ~13 μ s for data prep

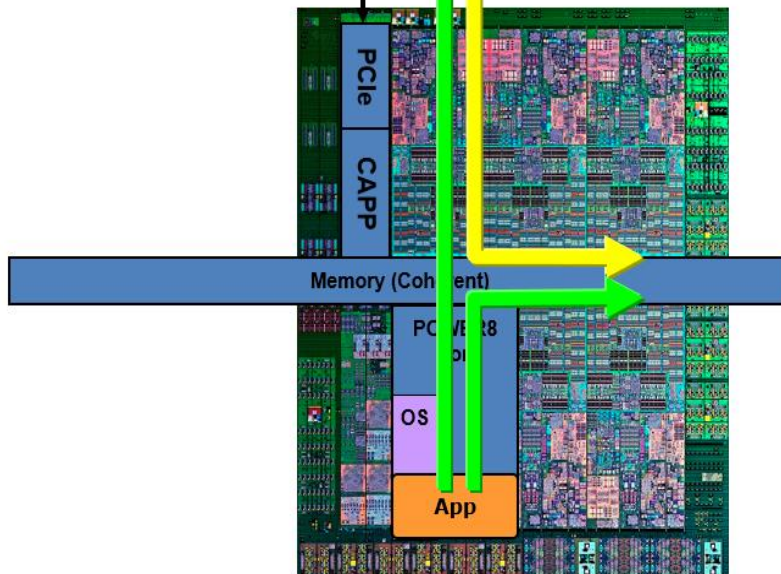
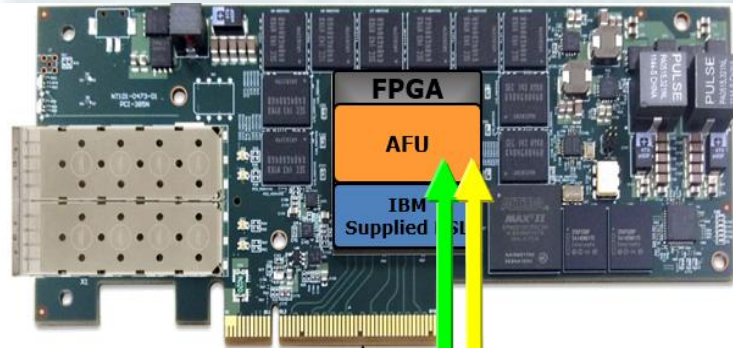


Flow with a Coherent Model: Total 0.36 μ s



IBM CAPI

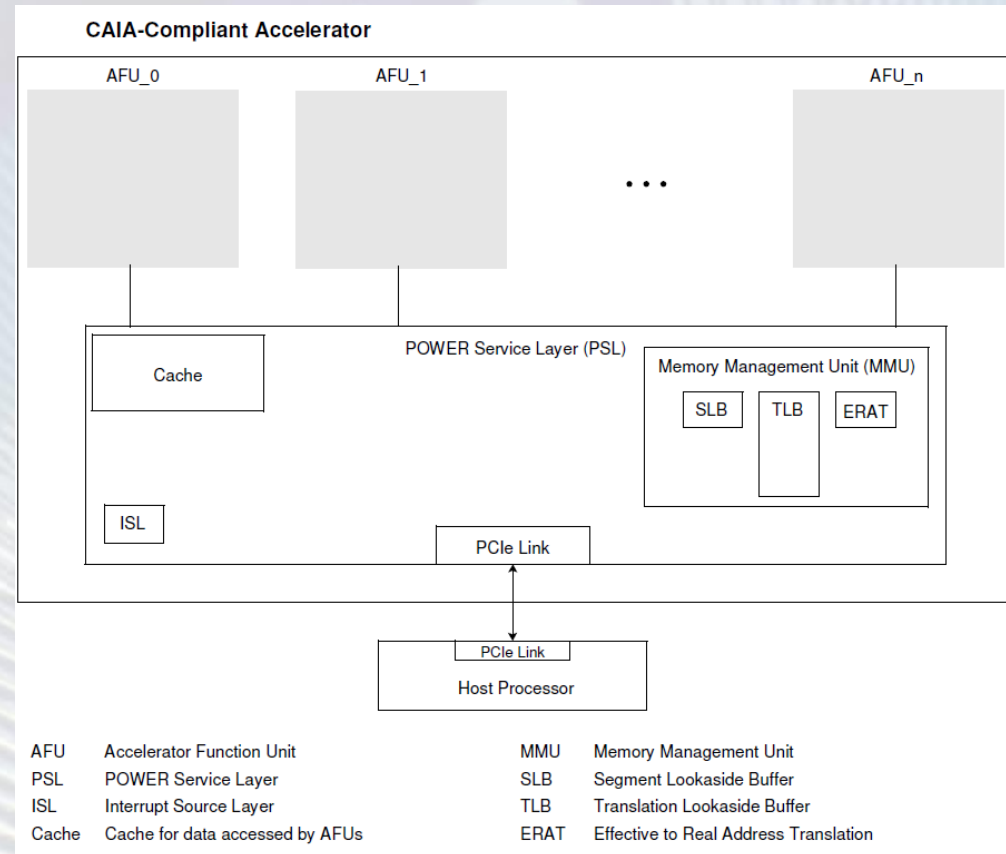
- A működésben részt vevő funkcionális egységek



- Proprietary hardware to enable coherent acceleration
 - Operating system enablement
 - Ubuntu LE
 - Libcxl function calls
 - Customer application and accelerator
- Application sets up data and calls the accelerator functional unit (AFU)
 - AFU reads and writes coherent data across the PCIe and communicates with the application
 - PSL cache holds coherent data for quick AFU access

IBM CAPI

- **CAIA Coherent Accelerator Interface Architecture**
 - Illeszkedik a CAPI Coherent Accelerator Processor Interface-hez
 - A lényeg a PSL egység, amiben CACHE, MMU, és ISL is van !
Ezt az IBM biztosítja
 - A CAIA egyidőben több Accelerator Functional Unit-ot (AFU) is tartalmazhat
 - Valódi dinamikus rekonfiguráció lehetősége

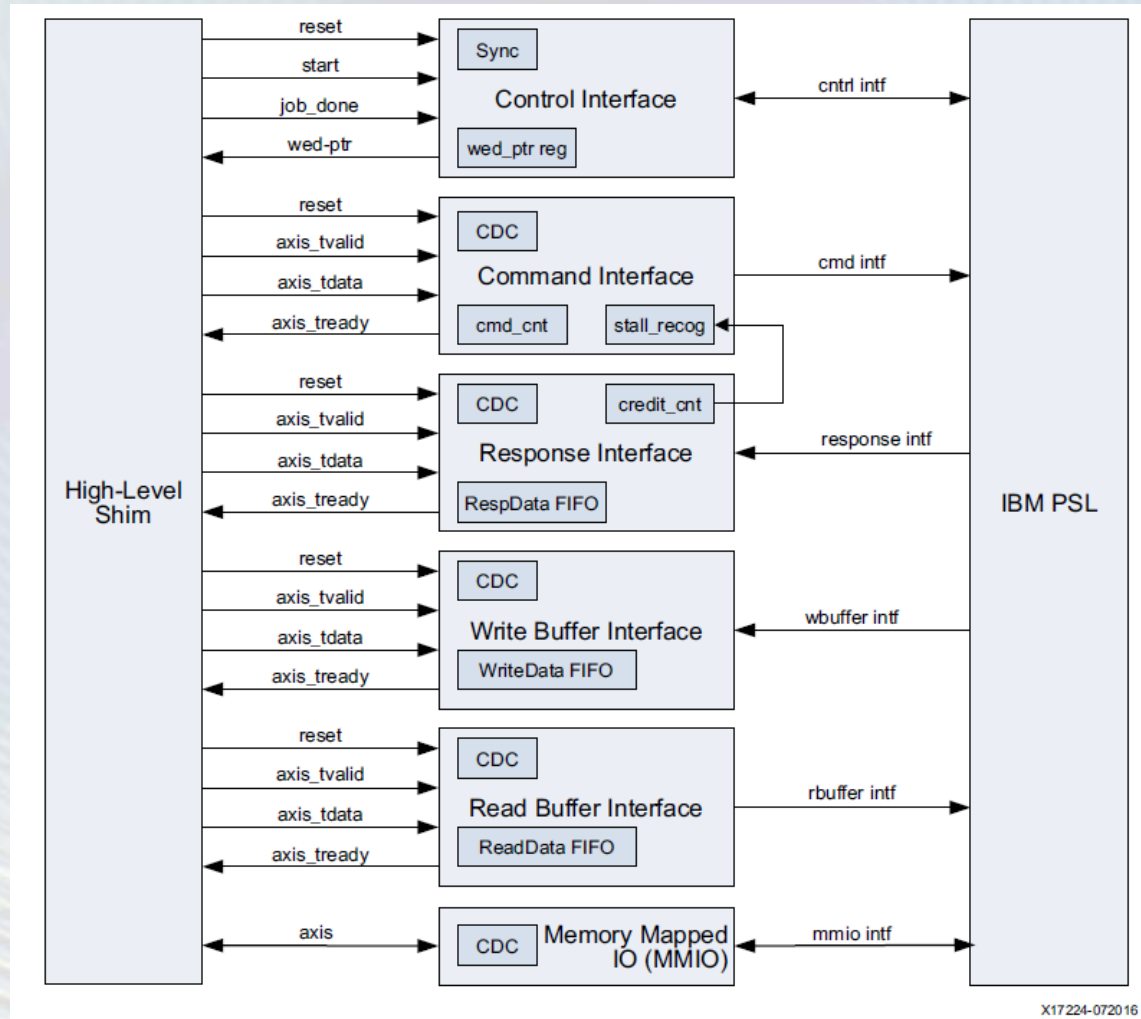


Xilinx IBM CAPI interfész

- **Kapcsolatot teremt a PSL és az AFU logika között**
- **Az FPGA-n belül a szabványos AXI interfészre épül**
- **A PSL 5 interfész csatornát használ:**
 - Vezérlő interfész
 - Parancs interfész
 - Válasz interfész
 - Írasi és olvasási adatbuffer interfészek
 - MMIO interfész (a gyorsító regisztereinek kezelésére)
 - Megjegyzés: az adatátvitel történhet out-of-order módon, az érvényes adatokat

Xilinx IBM CAPI interfész

- Az interfészek egy-egy AXI Streamre képeződnek
- Kivéve MMIO, ami AXI-Lite



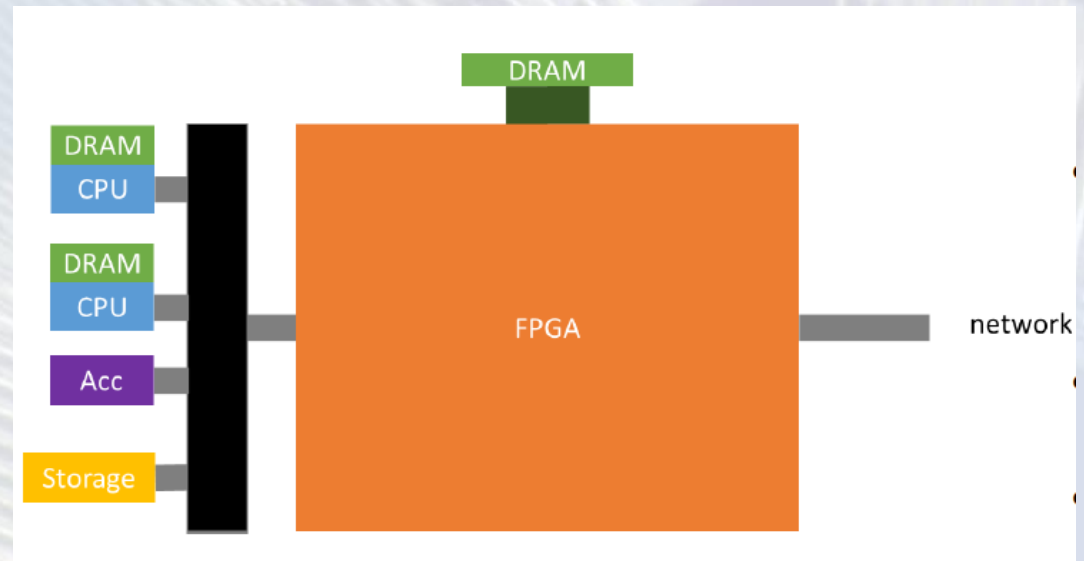
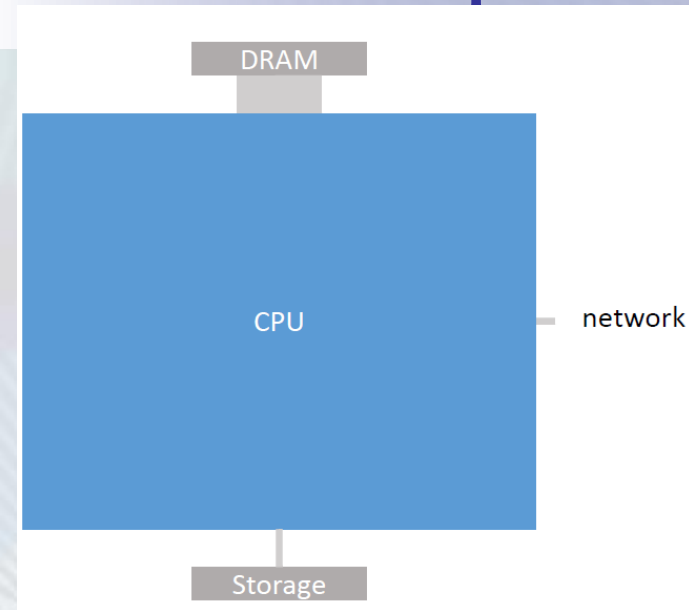
X17224-072016

CAPI továbblépés

- **CCIX Cache Coherent Interconnect for Accelerator**
- **Általános célú interfész specifikáció tetszőleges CPU + gyorsító közé**
- **Nyílt ipari szabvány kezdeményezés (AMD, ARM, Huawei, IBM, Mellanox, Xilinx)**
- **Nagy sávszélesség: 25Gb/s**
- **Bővítés memóriák (HMC, HBM) és tárolóeszközök (NVMe) felé**
- **Egyelőre kísérleti technológia**

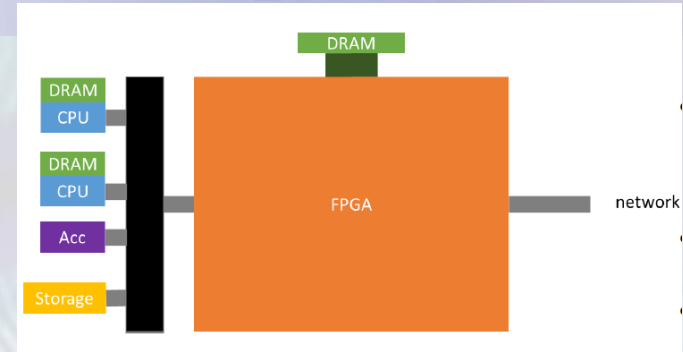
Microsoft Azure adatközpont

- Hagyományos számítógép modell
- Microsoft DataCenter Computer



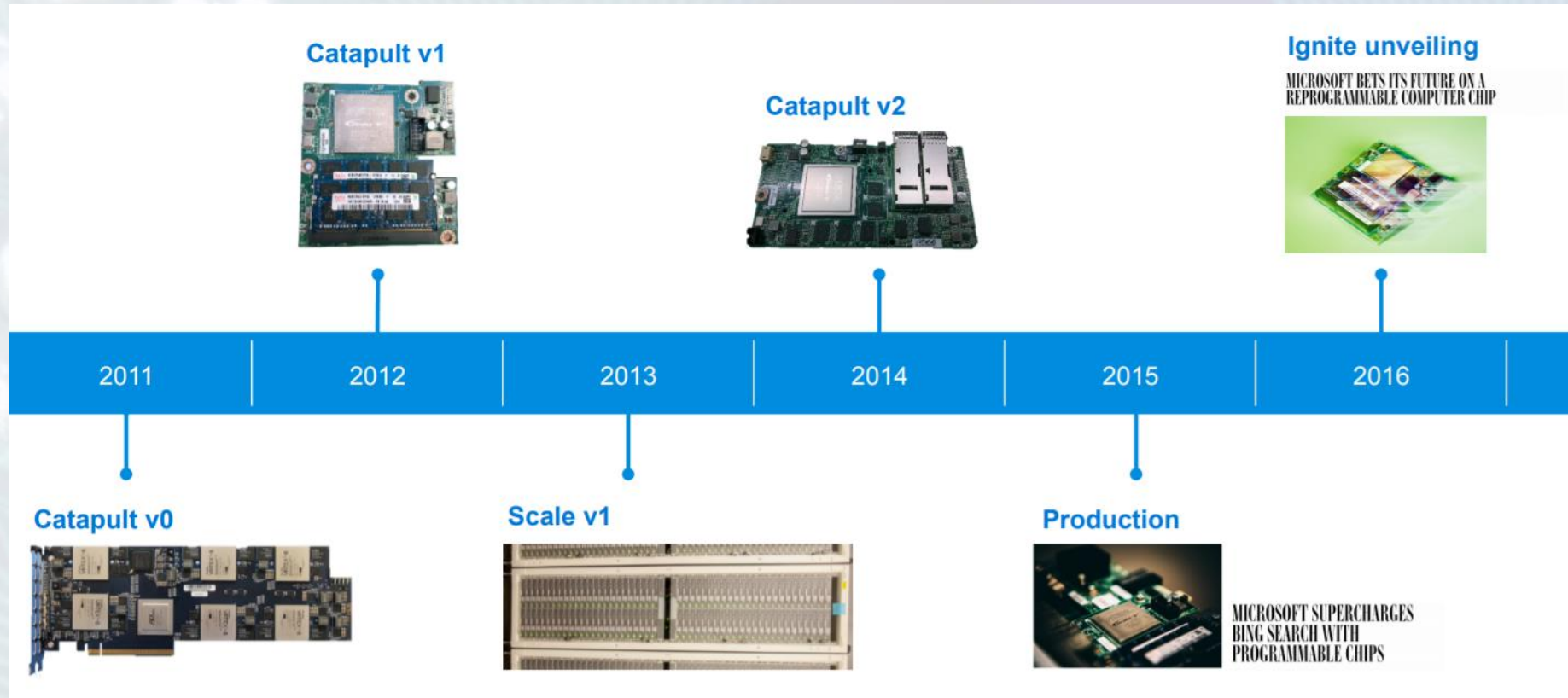
Microsoft Azure adatközpont

- Előnyök, tulajdonságok
- Az FPGA minden csomagot feldolgoz, „látja a forgalmat”
- Egyszerű feladatokat teljesen ellát
 - A csomagok a CPU-ba el sem jutnak
 - Közvetlenül kezeli a rendszermemóriát
- A CPU szerepe csak a nagykomplexitású feladatok ellátása
- Jellemző feladatok
 - Elosztott gépi tanulás
 - Software Defined Networking



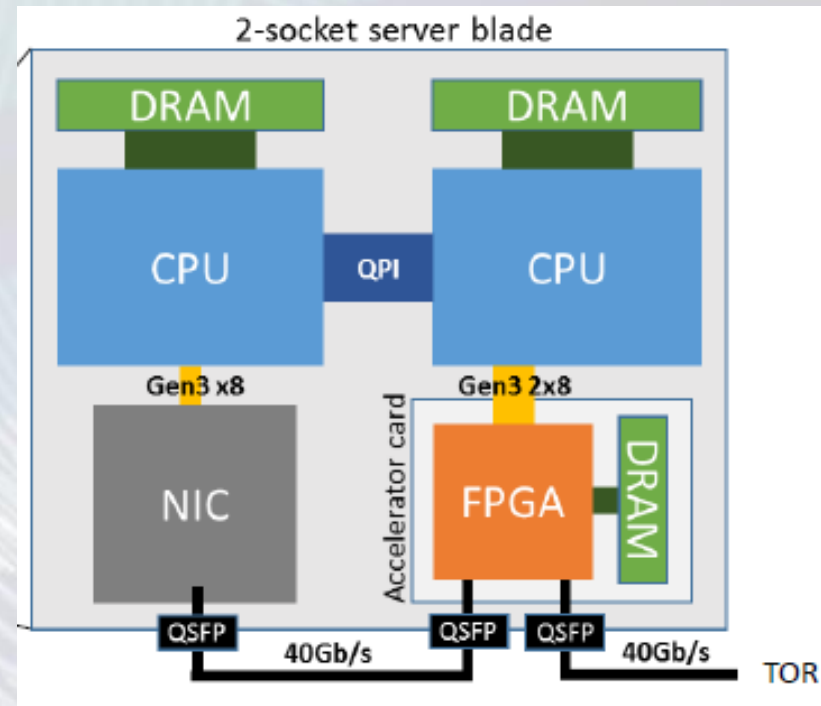
Microsoft Azure adatközpont

- A Catapult projekt eddigi fejlődése



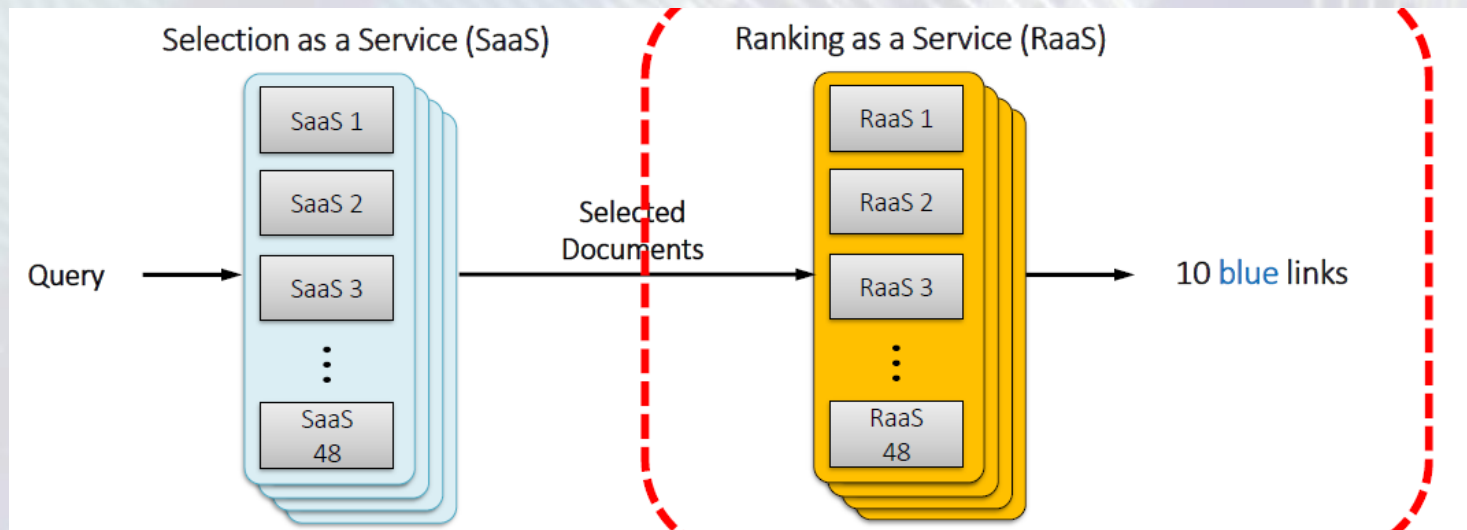
Microsoft Azure adatközpont

- A kialakított architektúra
- Teljesen flexibilis felépítés
 - Lokális gyorsítóeszköz
 - Távoli gyorsítóeszköz
 - Hálózati/tárolási gyorsító
- **ToR = Top-of-Rack** kommunikációs hálózat



Microsoft Azure adatközpont

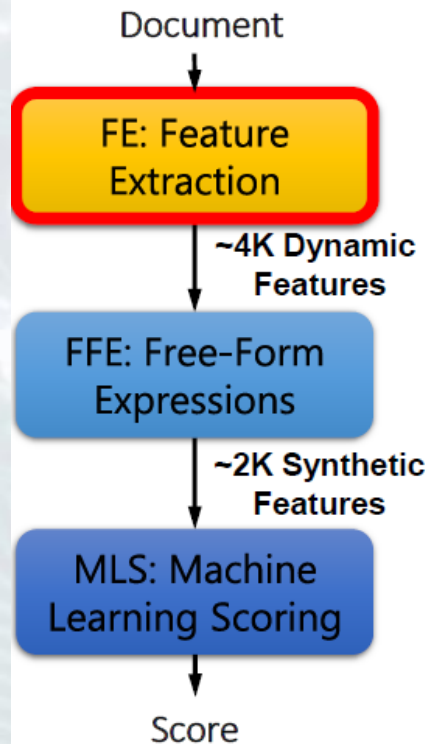
- **Lokális gyorsító alkalmazás**
 - A korábban már említett Intel CPU + FPGA megoldástól eltérően PCIe kapcsolat (a QPI foglalt)
 - A rendszermemória elérés PCIe-en keresztül
 - Bing keresésekhez, kiválasztás, sorbarendezés



Microsoft Azure adatközpont

FE: Feature Extraction

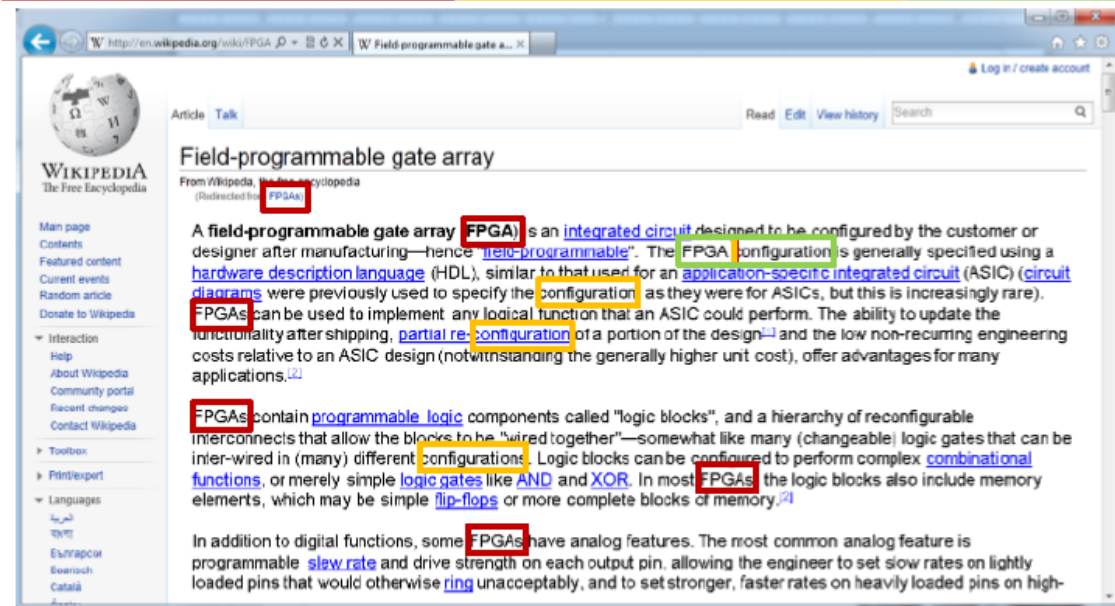
Query: "FPGA Configuration"



NumberOfOccurrences_0 = 7

NumberOfOccurrences_1 = 4

NumberOfTuples_0_1 = 1

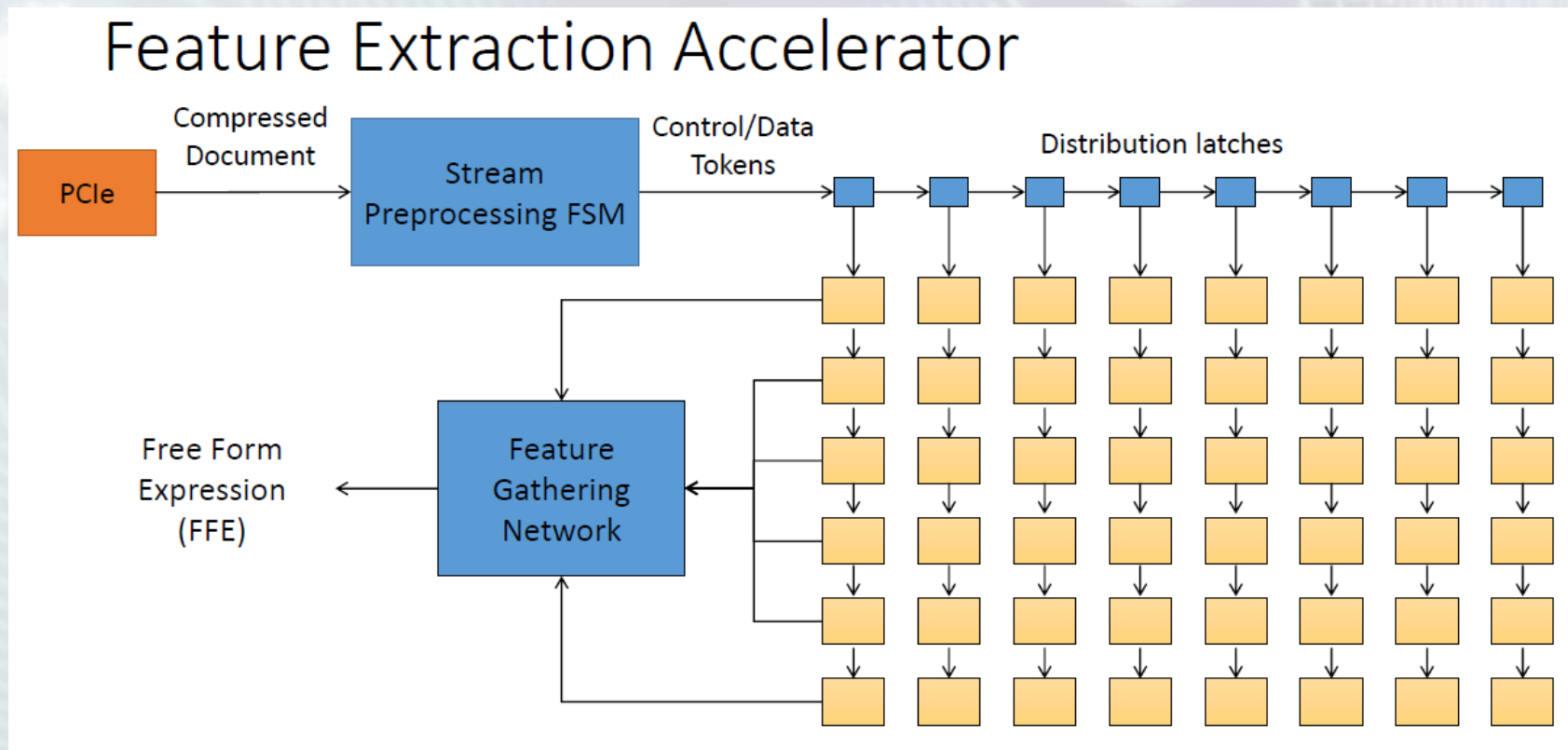


H2RC Nov 14, 2016

14

Microsoft Azure adatközpont

- Lehet egyetlen FPGA teljes belső konfigurációja, vagy az erőforrások egy része

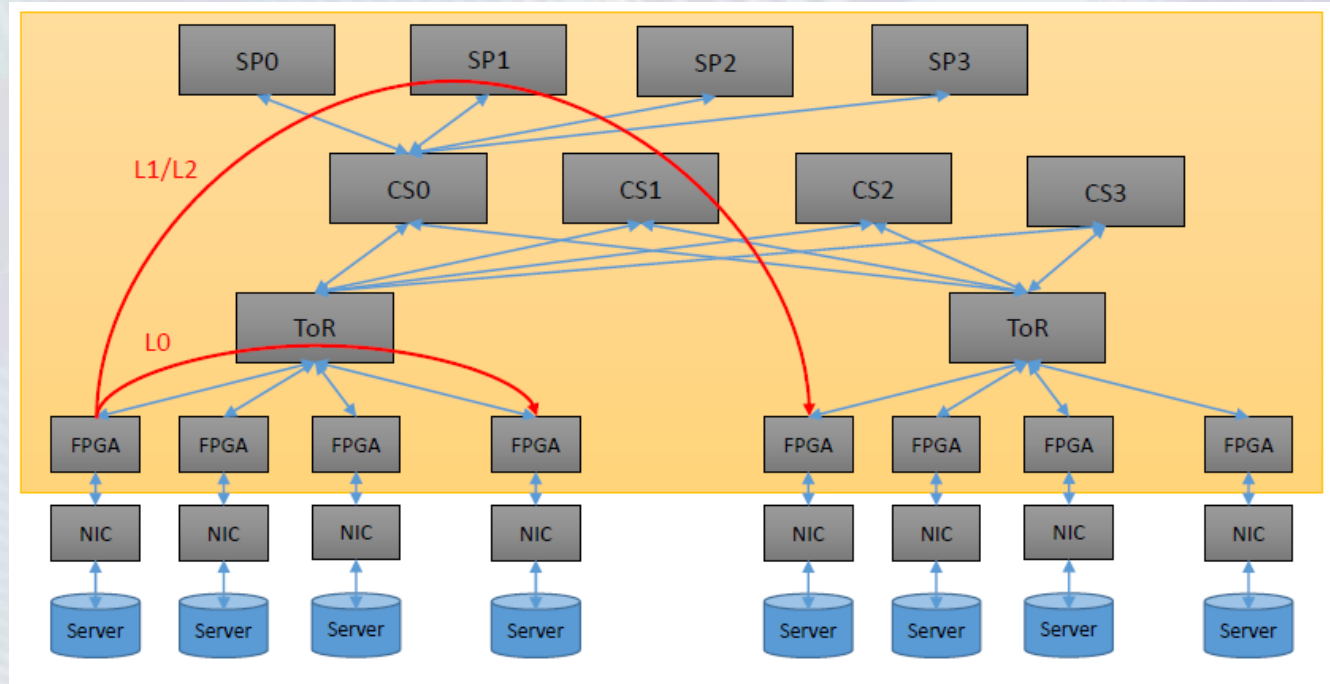


Microsoft Azure adatközpont

- **Távoli gyorsító alkalmazás**
 - A terhelések megvalósíthatósága nem mindig pontosan illeszkedik az FPGA méretéhez
 - A kihasználatlan területek elvesztegetett FPGA erőforrások vagy és elvesztegetett teljesítmény
 - Opció 1. Csökkentett teljesítmény/erőforrásigény, hogy elférjen egy FPGA-ba
 - Opció 2. Hozzáférés több szervertől hálózaton
 - Megoldás: ToR hálózati hozzáférés

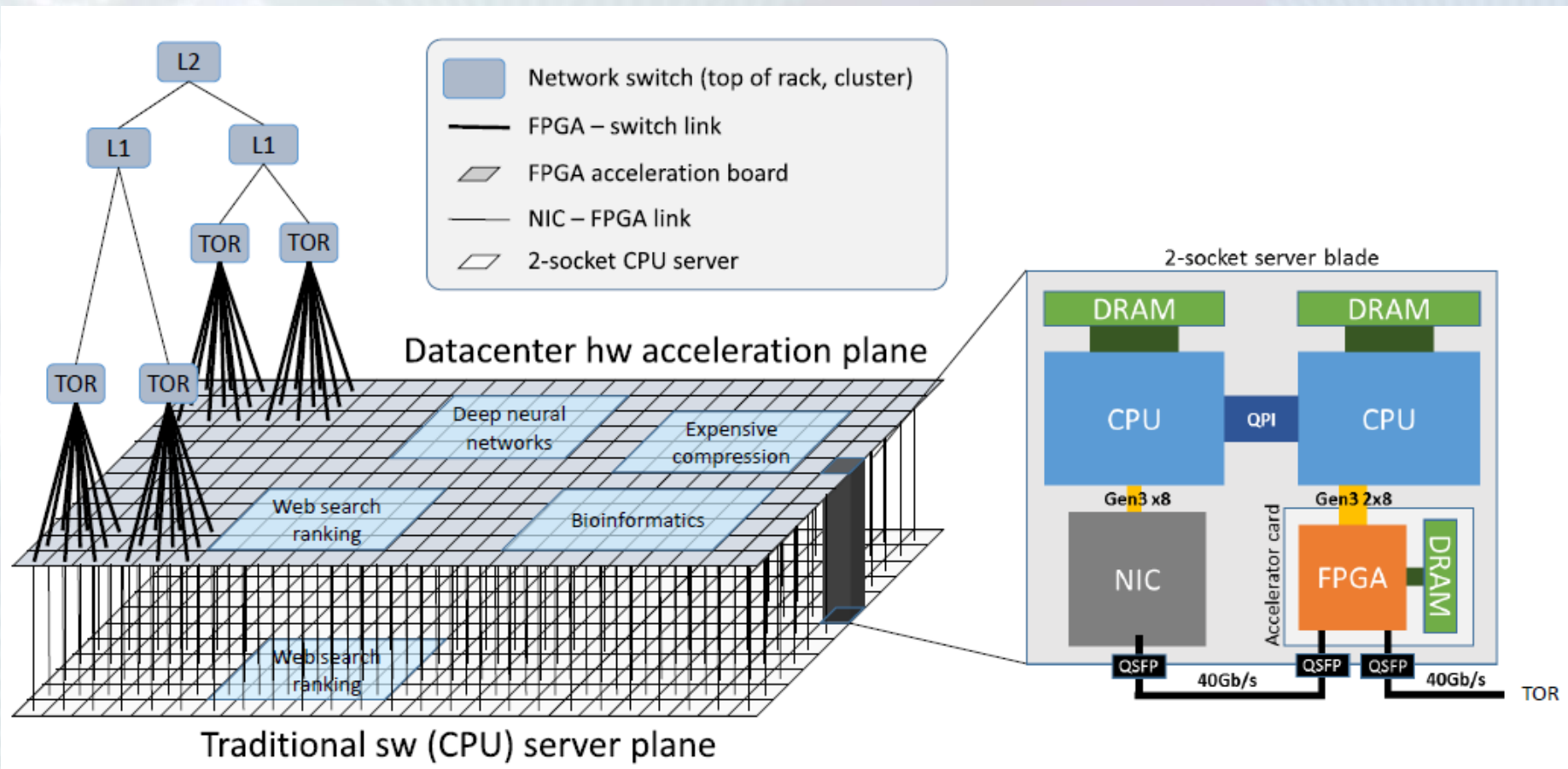
Microsoft Azure adatközpont

- FPGA-k közötti közvetlen hálózati kapcsolat
- Egy ToR alatt ($6 \times 8 = 48$ FPGA) közvetlen nagysebességű L0 szintű kapcsolat
- Ezek felett L1 és L2 szintű hálózati eszközök vannak



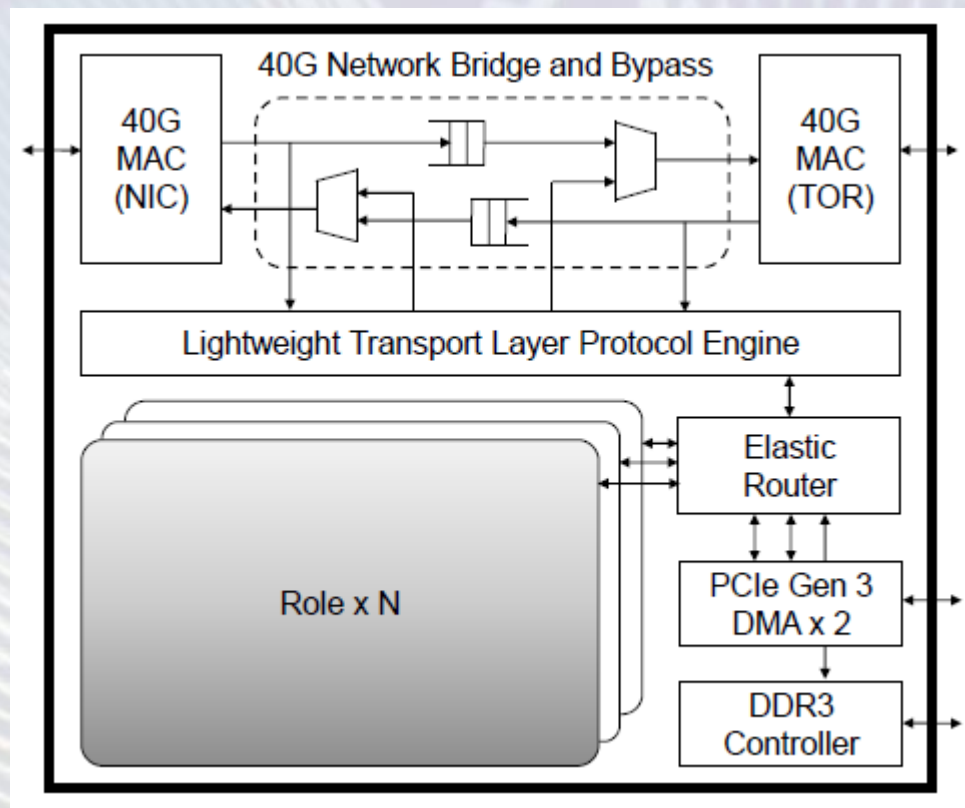
Microsoft Azure adatközpont

- **FPGA alapú konfigurálható felhő adatközpont**



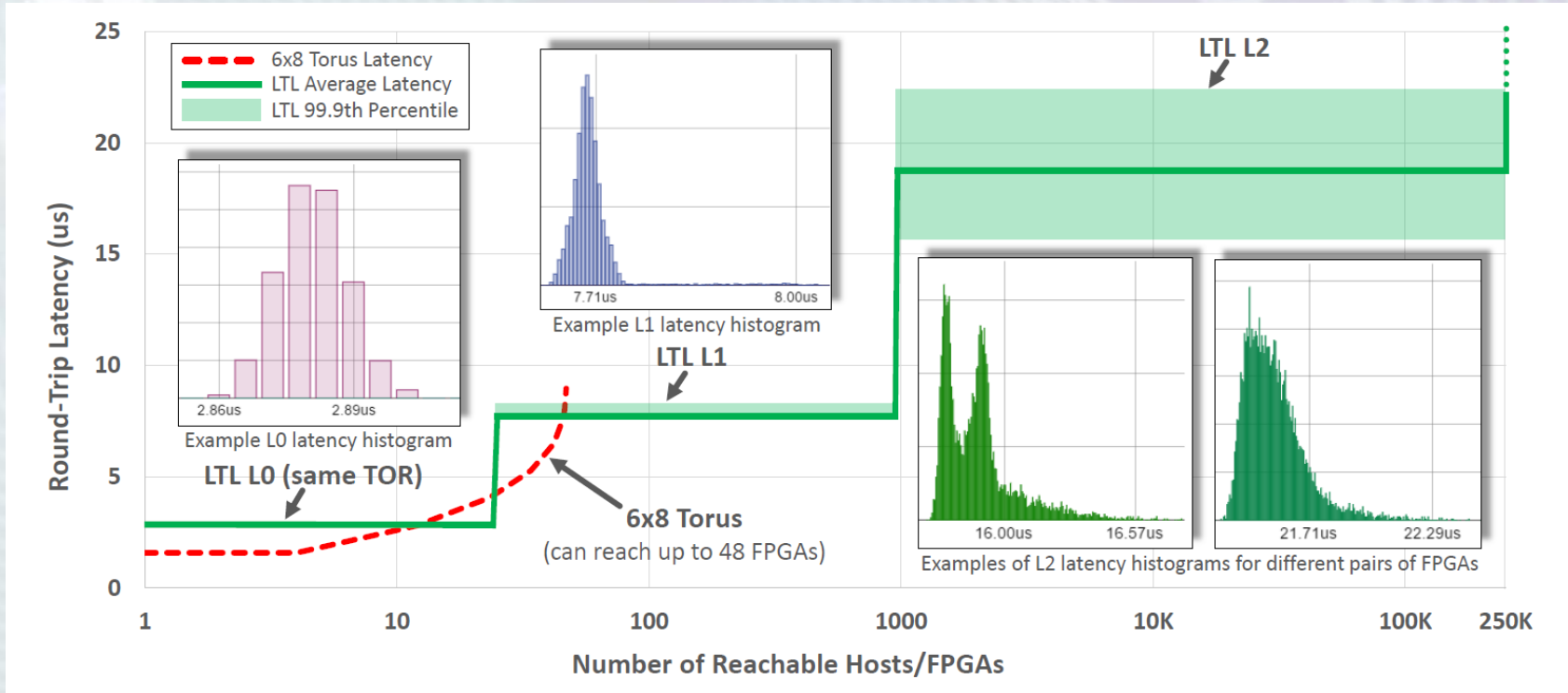
Microsoft Azure adatközpont

- Az FPGA-k belső felépítése
- Két komponens: Role, az alkalmazói logika, és a közös I/O és vezérlés
- Lokális memória 4GB
- Két 40Gb interfész
- PCIe IF DMA-val, közvetlen rendszer memória elérés
- Belső adatutak, belső hálózat



Microsoft Azure adatközpont

- **Lightweight Transport Layer Protocol Engine**
- **UDP és IP protokolt használ a hálózaton belül**
- **Bonyolult belső egység, megbízható adatátvitellel**

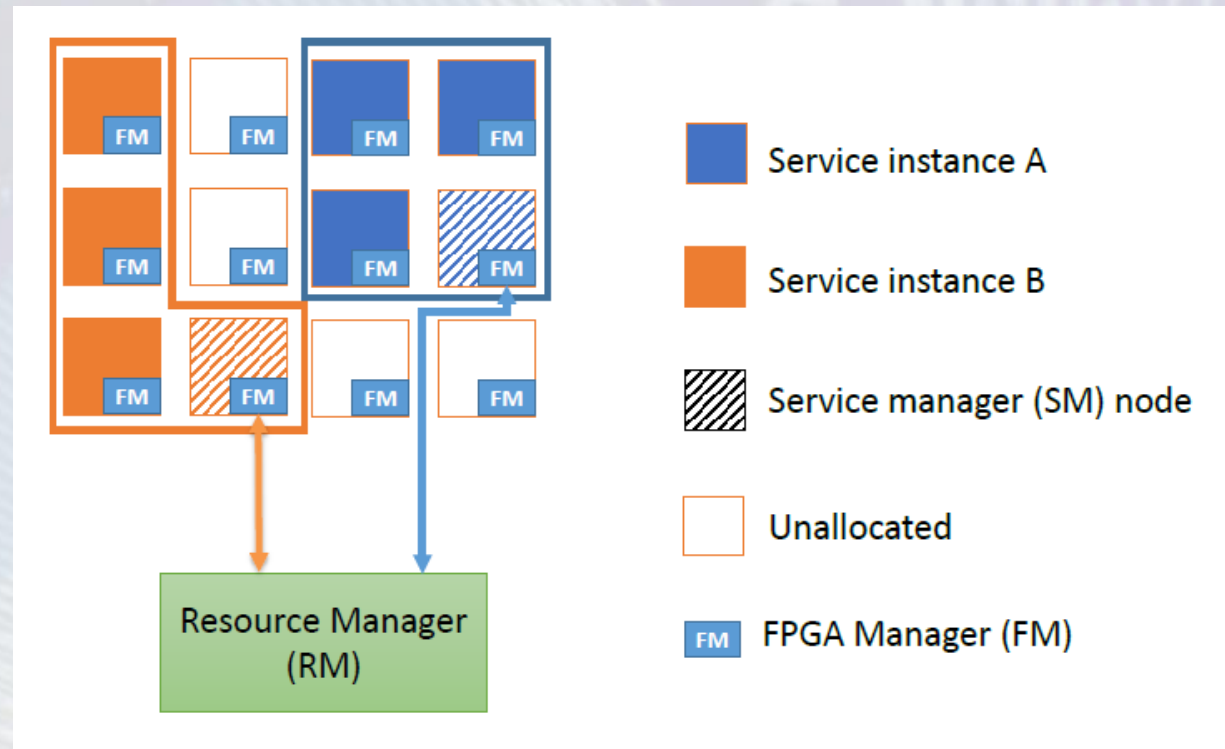


Microsoft Azure adatközpont

- **Lightweight Transport Layer Protocol Engine**
- **UDP és IP protokolt használ a hálózaton belül**
 - Bonyolult belső egység, megbízható adatátvitellel
- **Késleltetések**
 - L0 szinten, a ToR-on belül 2,88us (48 FPGA közvetlen elérése)
 - L1 szinten 7,7us, ezzel kb. 1000 FPGA érhető el
 - L2 szinten 16 us – 20 us, kb. 100 000+ FPGA felé
- **Valódi távoli és hálózati funkcionálisok**

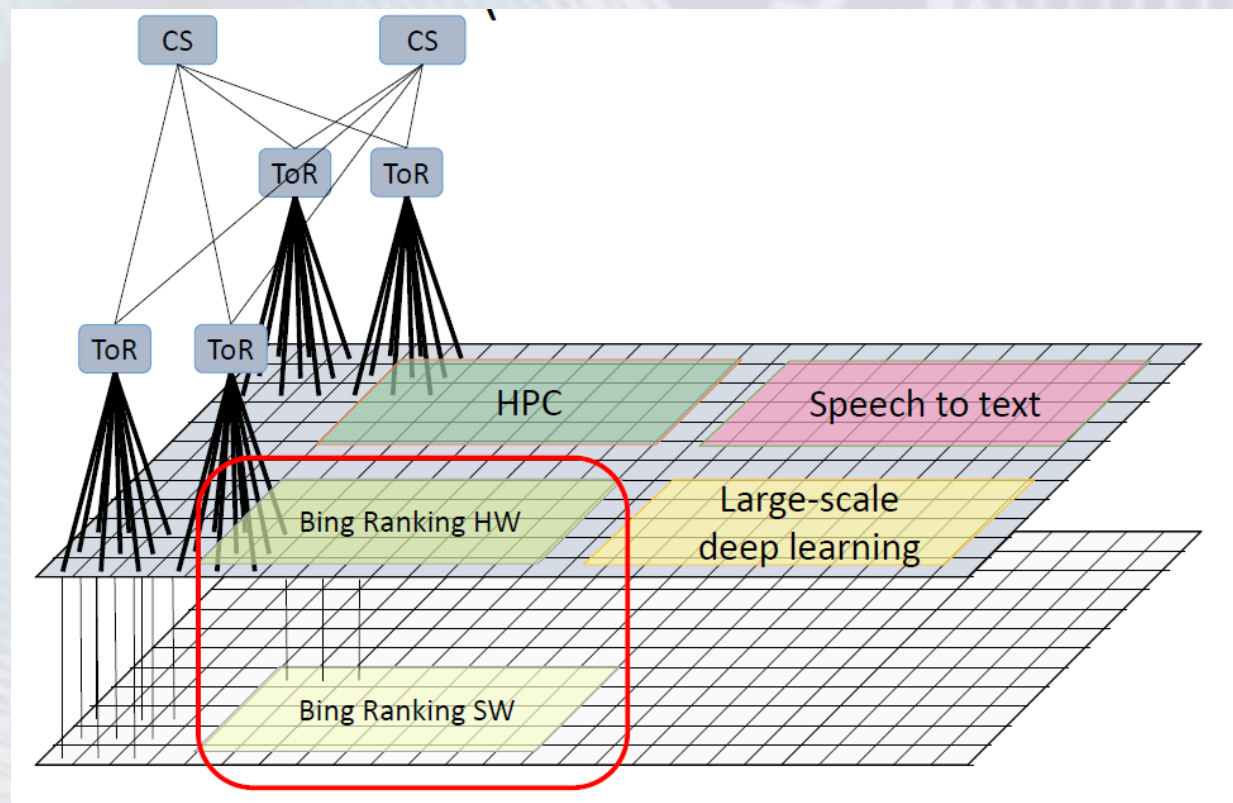
Microsoft Azure adatközpont

- Hardware-as-a-Service rendszer kialakítás
- Különböző szolgáltatások telepíthetők az FPGA-kra
- Egy-egy komponens egy vagy több FPGA-t használ
- Hierarchikus felügyelet
- RM Resource Manager
- SM Service Manager
- FM FPGA Manager



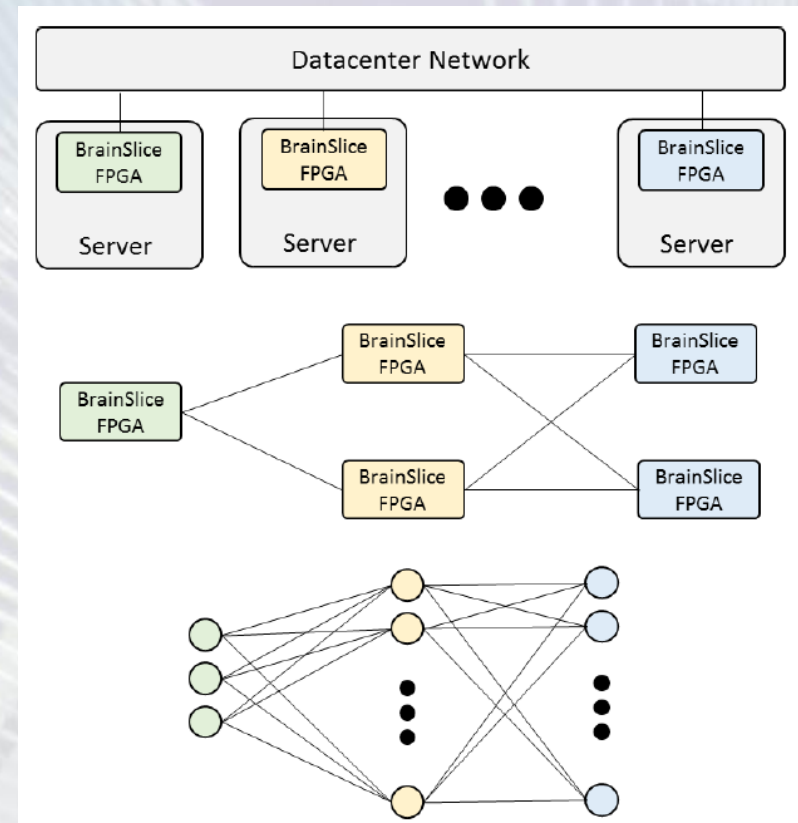
Microsoft Azure adatközpont

- **Hardware-as-a-Service rendszer kialakítás**
- **Különböző szolgáltatások telepíthetők az FPGA-kra**



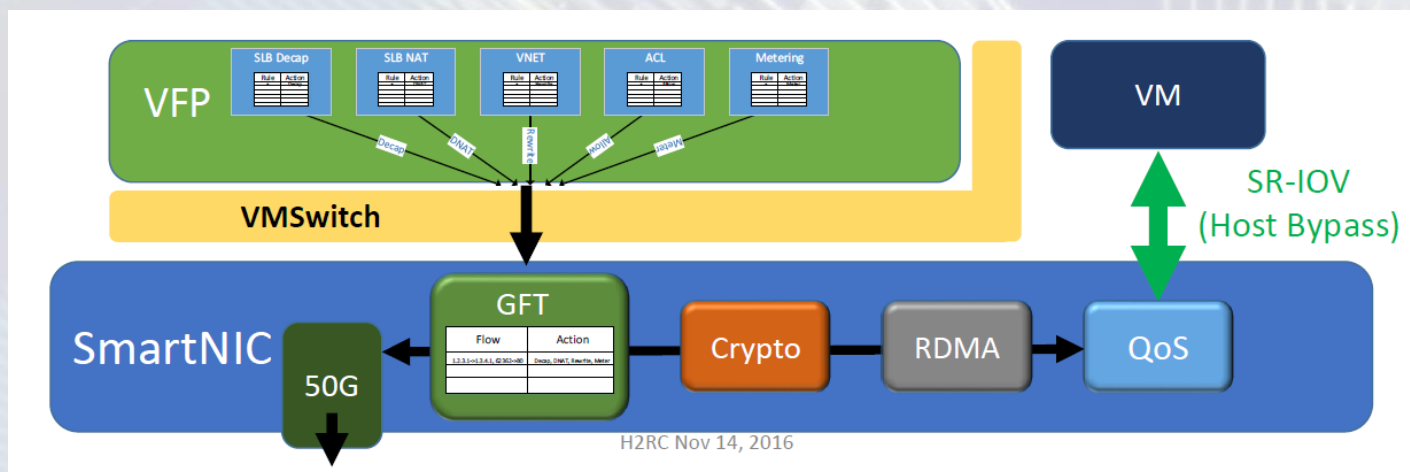
Microsoft Azure adatközpont

- **BrainWave: Extra méretű DNN hálózat**
- **Elosztott működés, akár 1000 FPGA között**
- **Optimális kapcsolati háló kapacitásra, késleltetésre, átviteli sávszélességre**



Microsoft Azure adatközpont

- Hálózati intelligens NIC felhő szolgáltatásokra
- Az Azure jelenleg a leggyorsabb publikált hálózat
 - 25Gb/s 25us késleltéssel
- Az FPGA minden csomagot kezel, a fejléc módosításával (sebessége gyakorlatilag a hálózati sebességgel azonos „bare metal”)



Összegzés

- **Az adatközpontok területe egy intenzív fejlesztési ág**
- **Előtérbe került a heterogán számítási rendszerek használata**
- **Célok: Energiahatékonyság, adatátviteli sebesség, minimális válaszidő késleltetés**
- **Mindez felhasználóbarát fejlesztési technológiákkal**
- **Az FPGA-k használata fejlődik, egyre jobban beépülnek a rendszerekbe**