

Újrakonfigurálható technológiák nagy teljesítményű alkalmazásai

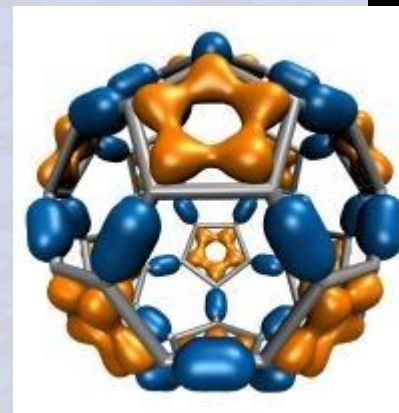
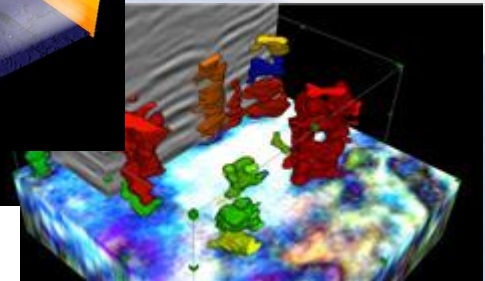
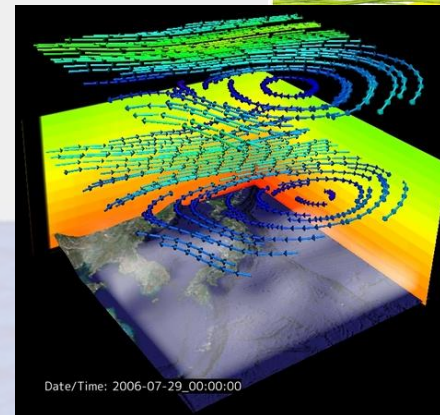
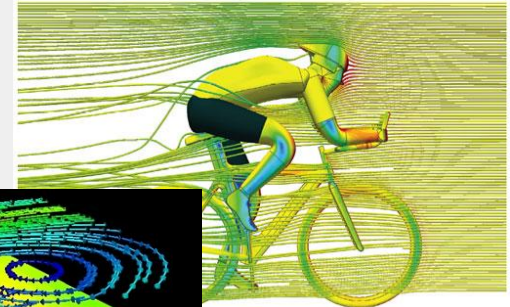
Architektúrák és kommunikációs interfészek

Szántó Péter

BME MIT, FPGA Laboratórium

HPC alkalmazások

- **Computational Fluid Dynamics (CFD)**
 - Folyadék (gáz) áramlás szimuláció
- **Időjárás előrejelzés**
- **Szeizmikus analízis**
 - Olaj/gáz kutatás
 - Földrengés előrejelzés
- **Kémiai informatika**
 - molekula fejlesztés
- **Bioinformatika**
 - Gén kutatás, dokkolás
- **Pénzügyi számítások**
 - Kockázat analízis



HPC TOP500

- Oak Ridge Jaguar – AMD Opteron Six Core 2.6 GHz
- IBM Roadrunner – AMD Opteron/PowerXCell
- Kraken XT5 – AMD Opteron
- JUGENE - Blue Gene/P, IBM PowerPC
- Tianhe-1 – Intel Xeon, ATI Radeon
- NASA Pleiades – SGI Altix ICE 8200EX, Intel Xeon
- IBM Blue Gene/L, IBM PowerPC
- IBM Blue Gene/P Solution, IBM PowerPC
- Ranger - SunBlade x6420, AMD Opteron
- Red Sky - Sun Blade x6275, Intel Xeon

HPC – TOP 500

- **Jaguar (Oak Ridge Lab)**

- AMD Opteron
- 224000 mag
- 2331 TFlops



- **IBM Roadrunner**

- Opteron/PowerXCell
- 122400 mag
- 1375 TFlops



HPC – TOP 500

- **IBM BlueGene/P**

- PowerPC 450
- 294912 mag
- 1002 Tflops



- **Tianhe-1**

- Intel Xeon
- 71680 mag
- 1375 TFlops
- 98304 GB memória



LRZ München – SGI Altix 4700

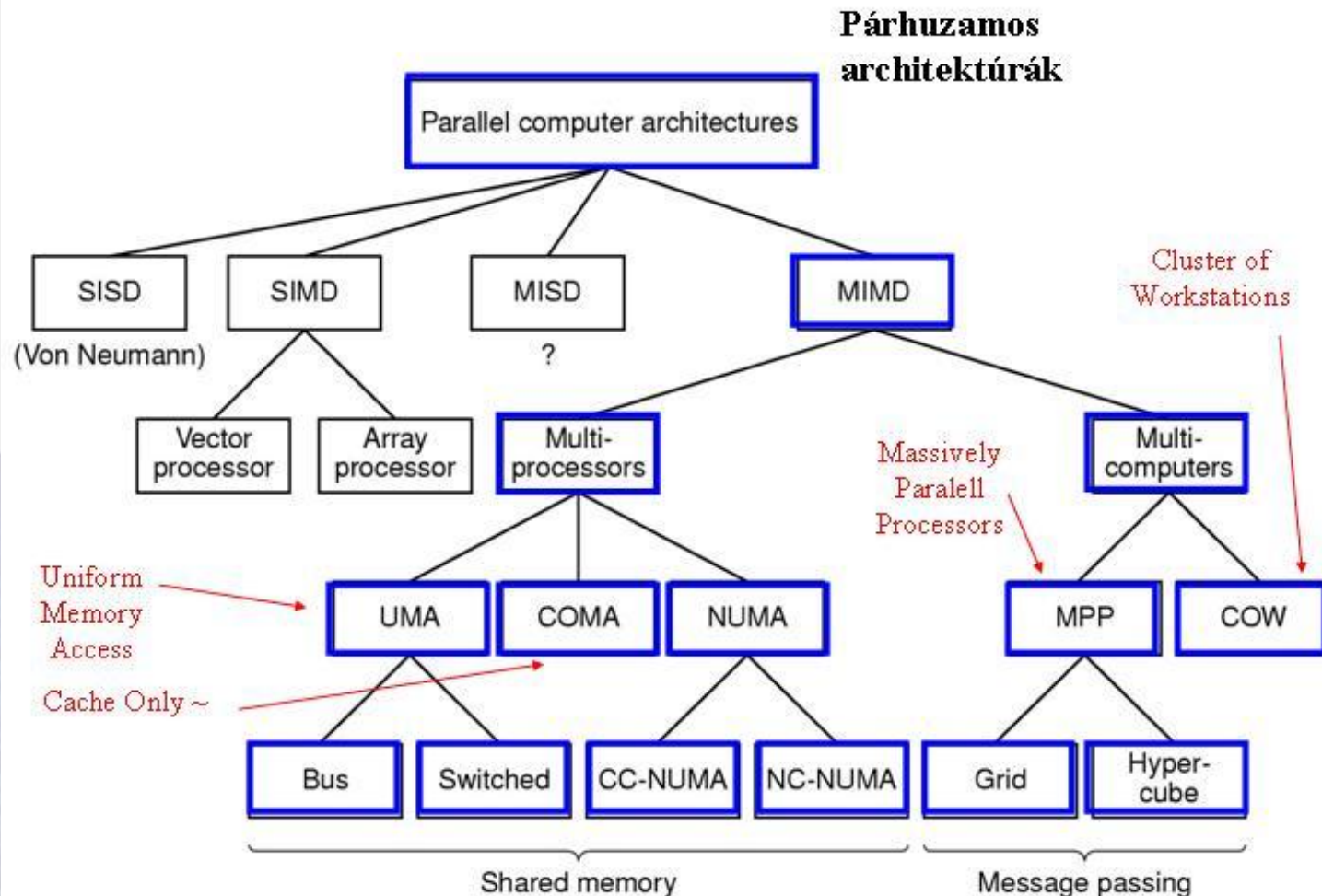


NASA Pleiades

51200 cores Altix ICE, 100 racks
608.83 Tflops RPeak Performance
51 TB of Memory [1GB/core]
storage SGI Nexis 9000, 3 Lustre cluster
two IB network 10D hypercube topology



Párhuzamos architektúrák



Alapfogalmak

- **Symmetric MultiProcessing**
 - Minden mag látja a teljes memóriát
 - Minden mag teljes hozzáféréssel rendelkezik a memóriához
 - Tipikusan max. 8 mag
- **Non-Uniform Memory Architecture (NUMA)**
 - Minden processzor saját memóriával rendelkezik, a processzoronkénti sávszélesség jóval nagyobb
 - A memória elérés függ a processzor – memória távolságtól
 - **ccNUMA**: cache coherent NUMA

Alapfogalmak

- **Pipeline végrehajtás**
- **Szuperskalár processzor**
 - Több utasítás fetch órajelenként (NEM == pipeline)
- **In-order utasítás végrehajtás**
 - Az assembly utasítások a kódnak megfelelő sorrendben hajtódnak végre (fetch – wait – execute - write)
- **Out-of-order (OoOE) utasítás végrehajtás**
 - A processzor átrendez(het)i az utasítások sorrendjét
 - Fetch - dispatch (wait) – execute – queue – write
- **Elágazás becslés**
 - Statikus (pl. backward – forward)
 - Dinamikus (pl. szaturációs számláló, adaptív prediktor)

Alapfogalmak

- **SISD**

- Single Instruction Single Data

- **SIMD**

- Single Instruction Multiple Data
- Intel MMX, SSEx, AMD 3DNow!, Intel AVX, Power AltiVec, ARM NEON

- **MIMD**

- Multiple Instruction Multiple Data
- VLIW: Very Long Instruction Word
 - Utasítás szintű párhuzamosítás fordítási időben

Alapfogalmak (MT)

- **Többszálúság (multi-threading)**
 - **Temporal MT:** az egymást követő utasítások származhatnak más szálból
 - **Simultaneous MT:** a processzor feldolgozó egységei ugyanazon órajelben más-más szállal foglalkoznak (szuperskalár)
 - **Multi-core:** több, teljes processzor mag egy szilíciumon
 - Intel HT: SMT, 2 szál
 - IBM POWER5, POWER6: SMT, 2 szál
 - IBM POWER7: SMT, 4 szál
 - Sun Niagara: TMT, 8 szál

Intel Xeon 5400/7400

- **Intel Core mikroarchitektúrájú processzorok**

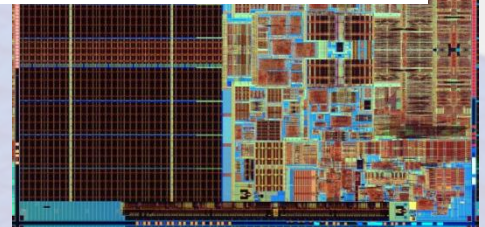
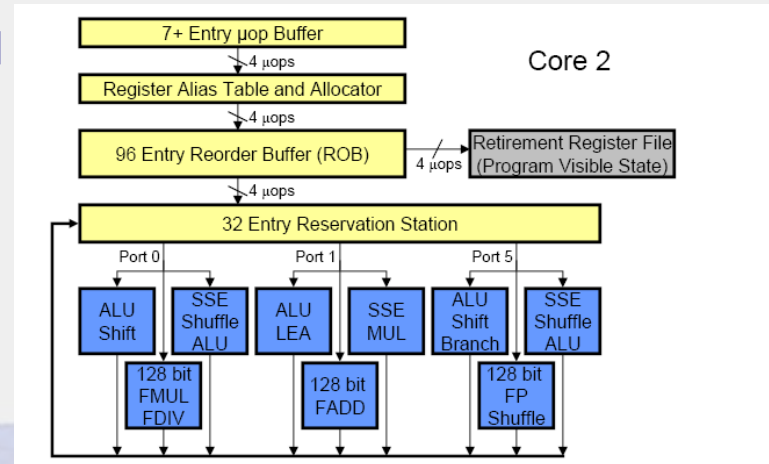
- 14 lépcsős pipeline
- ALU: 3 végrehajtóegység
- FPU: FMUL/FDIV + FADD
- SSE: 3x128 bit végrehajtóegység

- **Xeon 5400, Socket 771**

- Max: FSB 1600 MHz, 4 mag, 3,4 GHz, \$1500
- 256 kbyte L1, 2x6 mbyte L2 cache

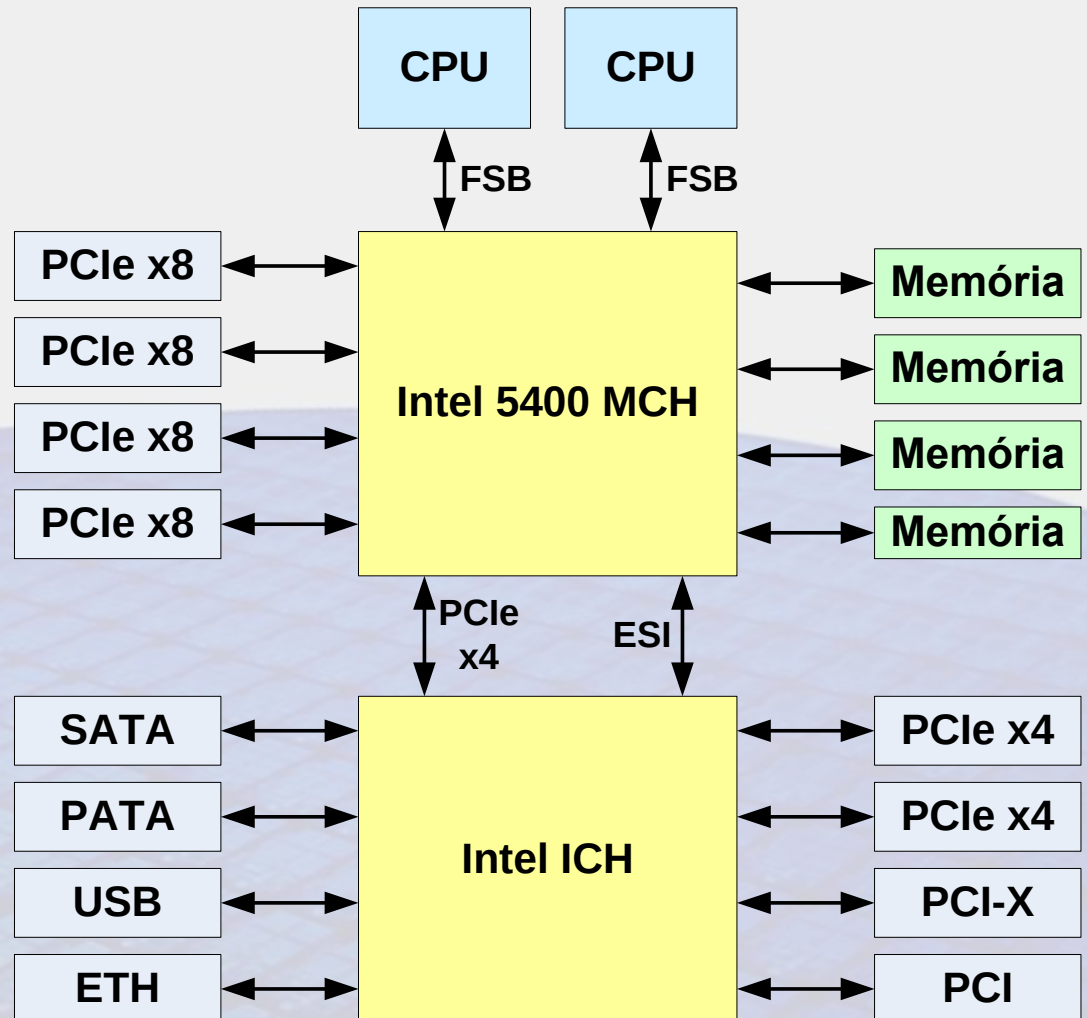
- **Xeon 7400, PGA604, 2mrd tr.**

- Max: FSB 1066 MHz, 6 mag, 2,66 GHz, \$2800
- 256 kbyte L1, 3x3 mbyte L2, 12 mbyte L3 cache



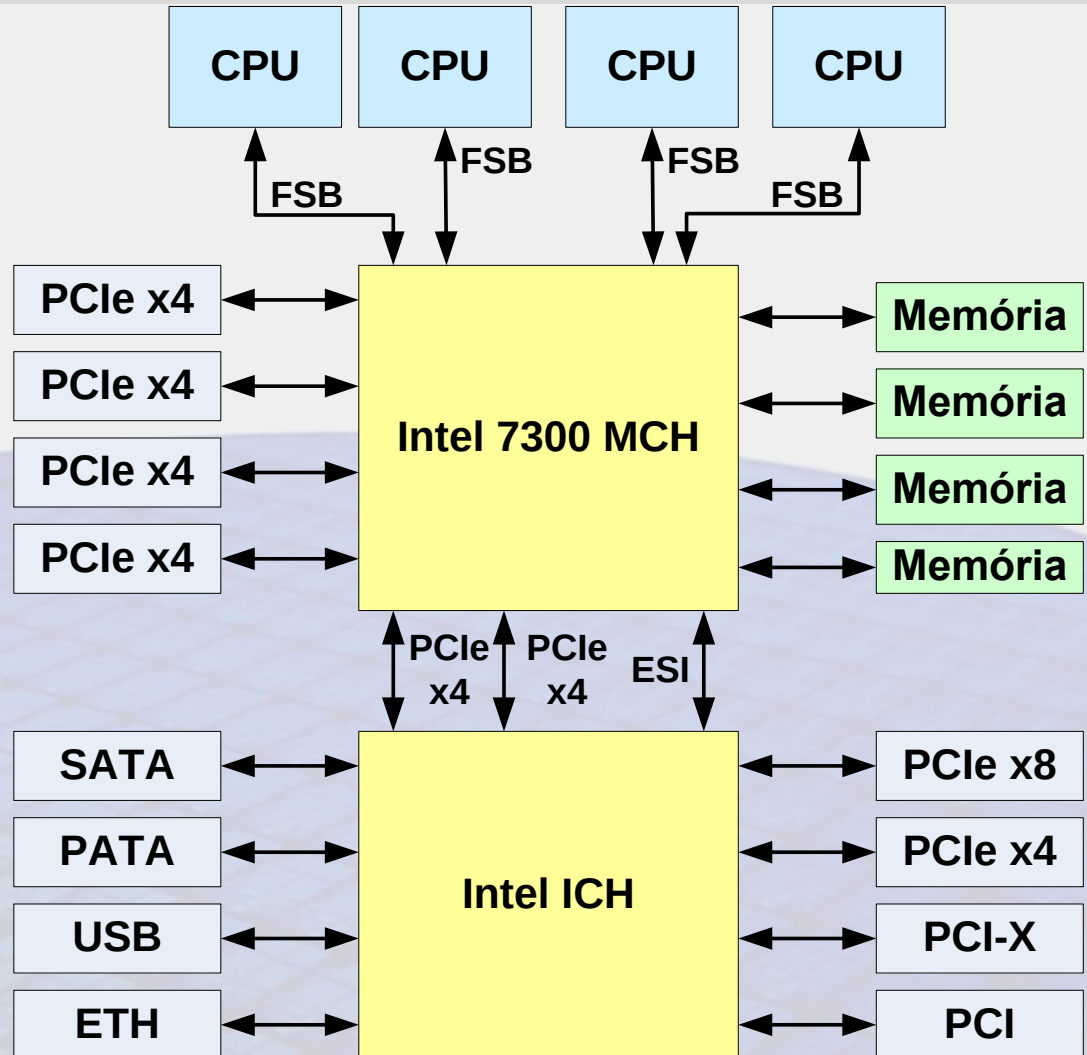
Xeon 5400

- **2 utas**
 - LGA771 foglalat
- **Memória**
 - DDR2-667
 - 4x64 bit
 - Max. 16 DIMM
 - Max. 128 Gbyte
- **PCIe 1.1**
 - Max. x16
- **PCI-X**
 - 133 MHz



Xeon 7400

- **4 utas**
 - PGA604 foglalat
- **Memória**
 - DDR2-667
 - 4x64 bit
 - Max. 32 DIMM
 - Max. 512 Gbyte
- **PCIe 1.1**
 - Max. x8
- **PCI-X**
 - 133 MHz

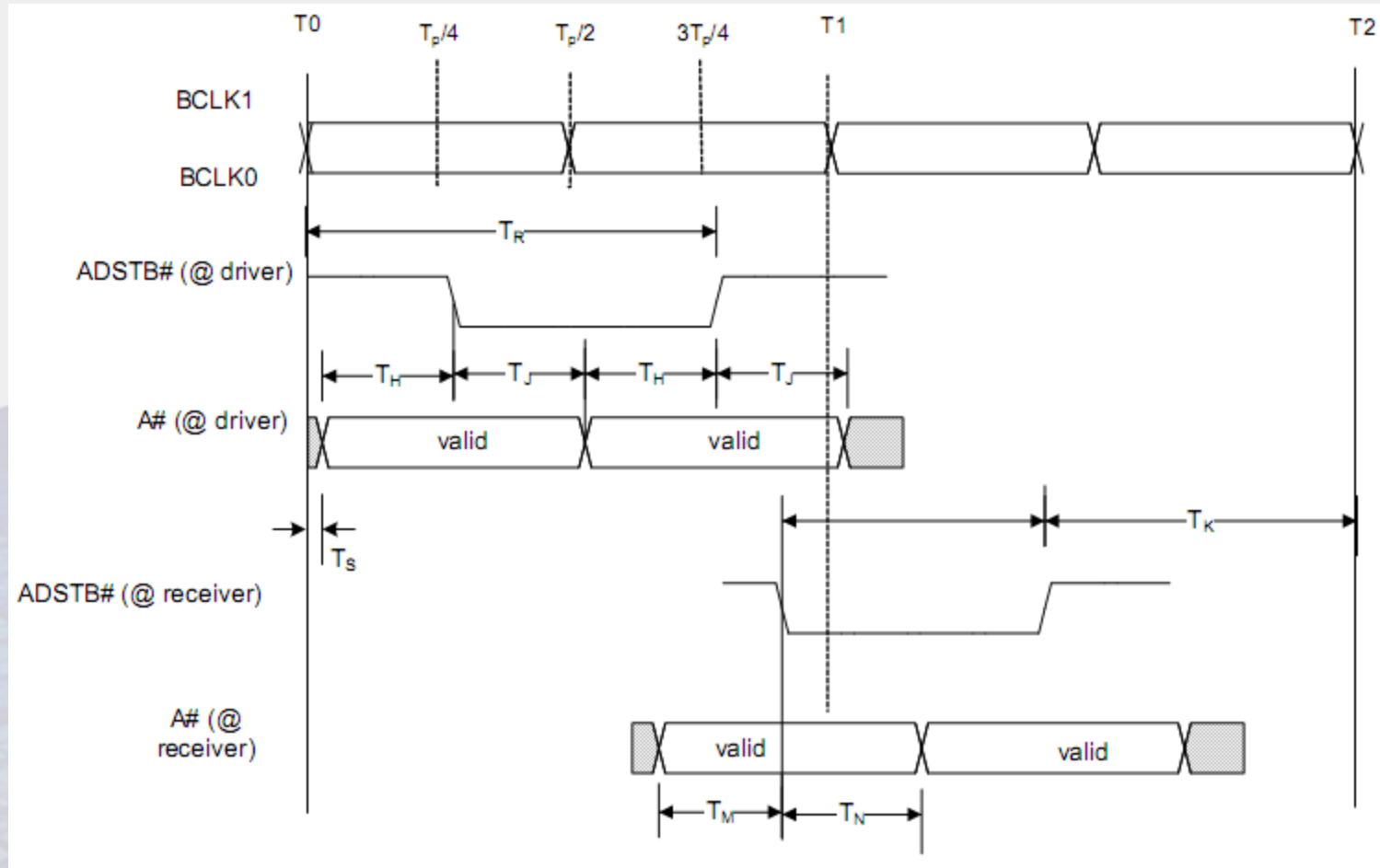


Front Side Bus (FSB)

- **„Igazi” párhuzamos busz (~170 vezeték)**
 - Szétválasztott cím és adatbusz
 - 36 bites címbusz
 - 2 irányú, 64 bites adatbusz, strobe jelek 2 byte-onként
 - Max. 5 egység (800 MHz) – max. 2 egység (1066 MHz)
 - „quad data rate”
- **Minden kommunikáció az FSB-n keresztül!**
 - utasítások beolvasása a központi memóriából
 - adatok beolvasása a memóriából, ill. kiírása a központi memóriába
 - adatcsere a processzorok között
 - periféria hozzáférések (a periféria buszától függetlenül)

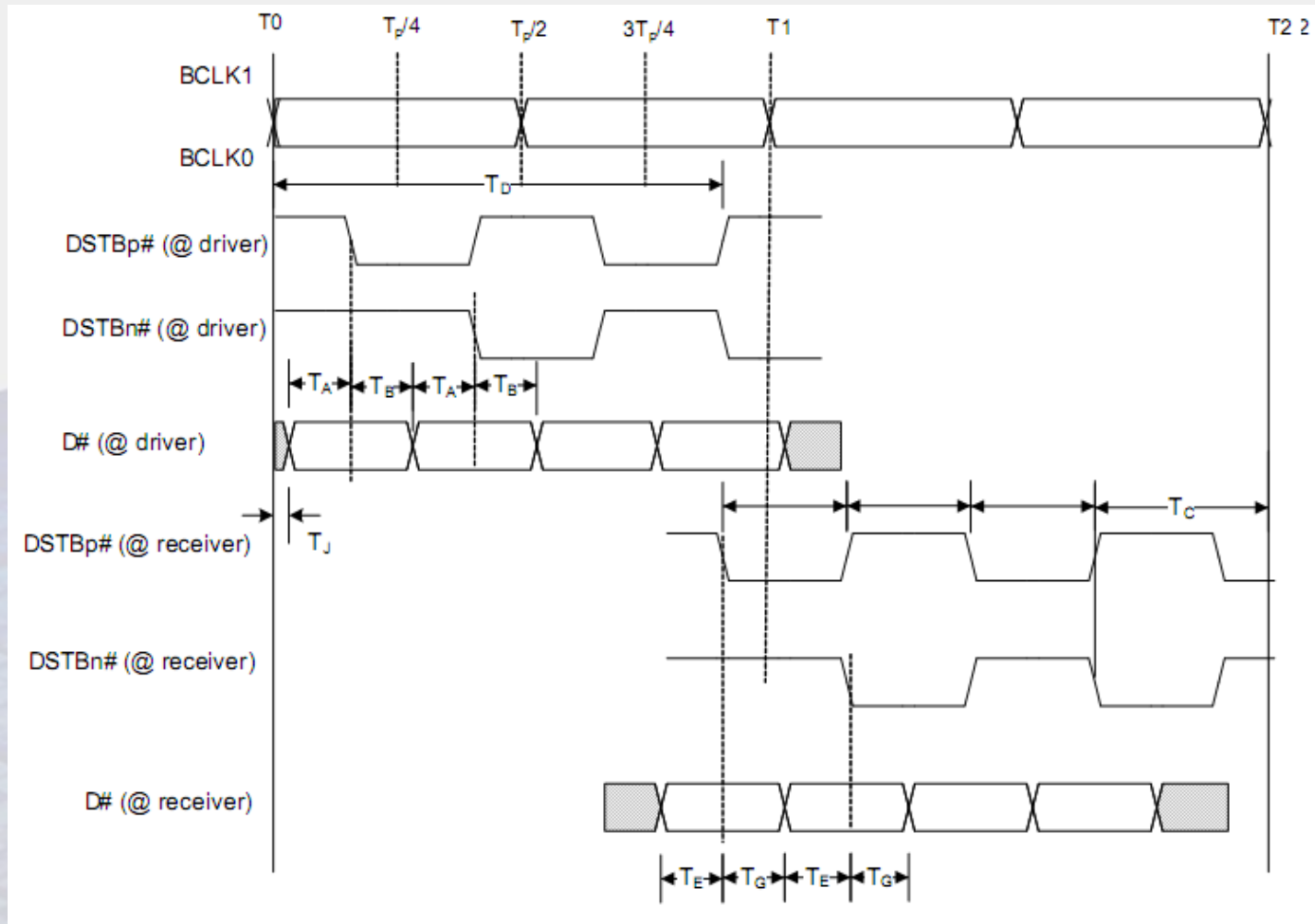
FSB cím/control ciklus

- Cím/control: double data rate



FSB adatciklus

- Adat: quad data rate



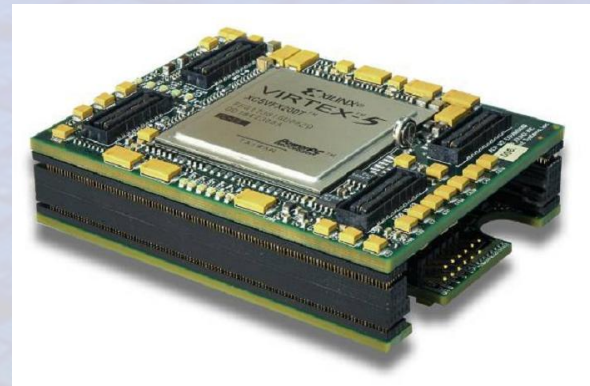
XtremeData XD2000i

- LGA771 foglalatba illeszkedő FPGA alapú gyorsító
 - Max. 1066 MHz FSB
- 3 Altera Stratix-III FPGA
 - 1 FSB interfész
 - 2 szabadon felhasználható
 - 9,6 Gbyte/s busz kapcsolat
 - 142k LE/FPGA
 - 5,5 Mbit BRAM, 378 MUL
- 2x8 Mbyte QDRII SRAM
 - 2x2,8 Gbyte/s



Nallatech FSB Accelerator

- **PGA604 foglalatba illeszkedő FPGA alapú gyorsító**
 - Max. 1066 MHz FSB, stack-struktúra
- **1x FSB interfész modul, busz kapcsolat a többi modullal**
- **Felhasználói modulok**
 - 2x LX330T/SX240T/FX200T FPGA
 - ~max 2 x 33000 LE
 - 2x2 bank DDR-II SRAM
 - Közvetlen FPGA-FPGA
 - 1x LX330T/SX240T/FX200T FPGA
 - ~max 33000 LE
 - 4 bank DDR-II SRAM
 - 20 GTP I/O



AMD Opteron

- **AMD K10 mikroarchitektúra**

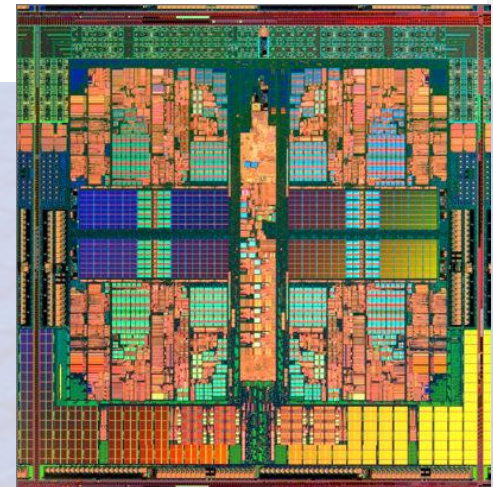
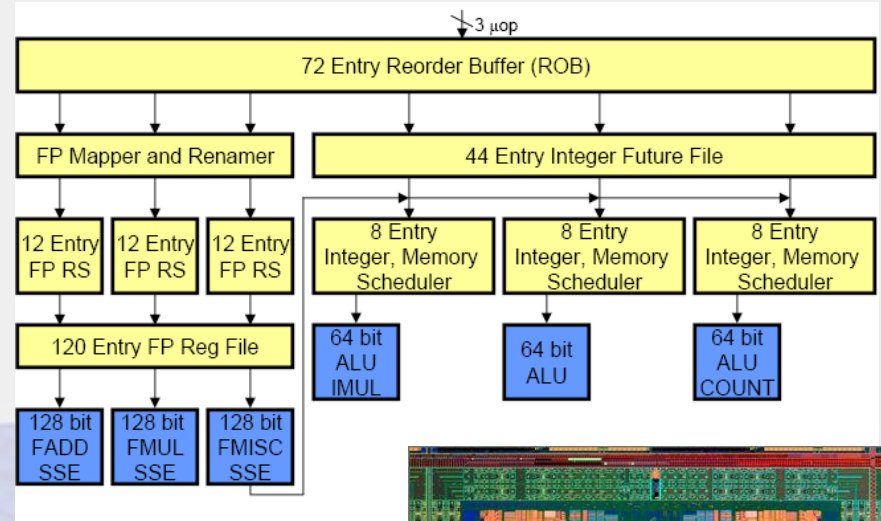
- 12 lépcsős pipeline
- ALU: 3 végrehajtóegység
- SSE/FPU: 3 végrehajtóegység
- Integrált DDR2 memória vezérlő (800 MHz)

- **AMD Opteron 23xx/837x (700m tr.)**

- 4 mag; 3,1 GHz; \$870/\$2600
- 4x64k L1, 4x512k L2, 6M L3 cache

- **AMD Opteron 24xx/84xx (900m tr.)**

- 6 mag; 2,8 GHz; \$1000/\$2600
- 6x64k L1, 6x512k L2, 6M L3 cache



AMD Opteron

- A processzorok HyperTransport linken keresztül kapcsolódnak (a chipszethez is)

- x3xx: 1,1 GHz
- x4xx: 2,2 GHz

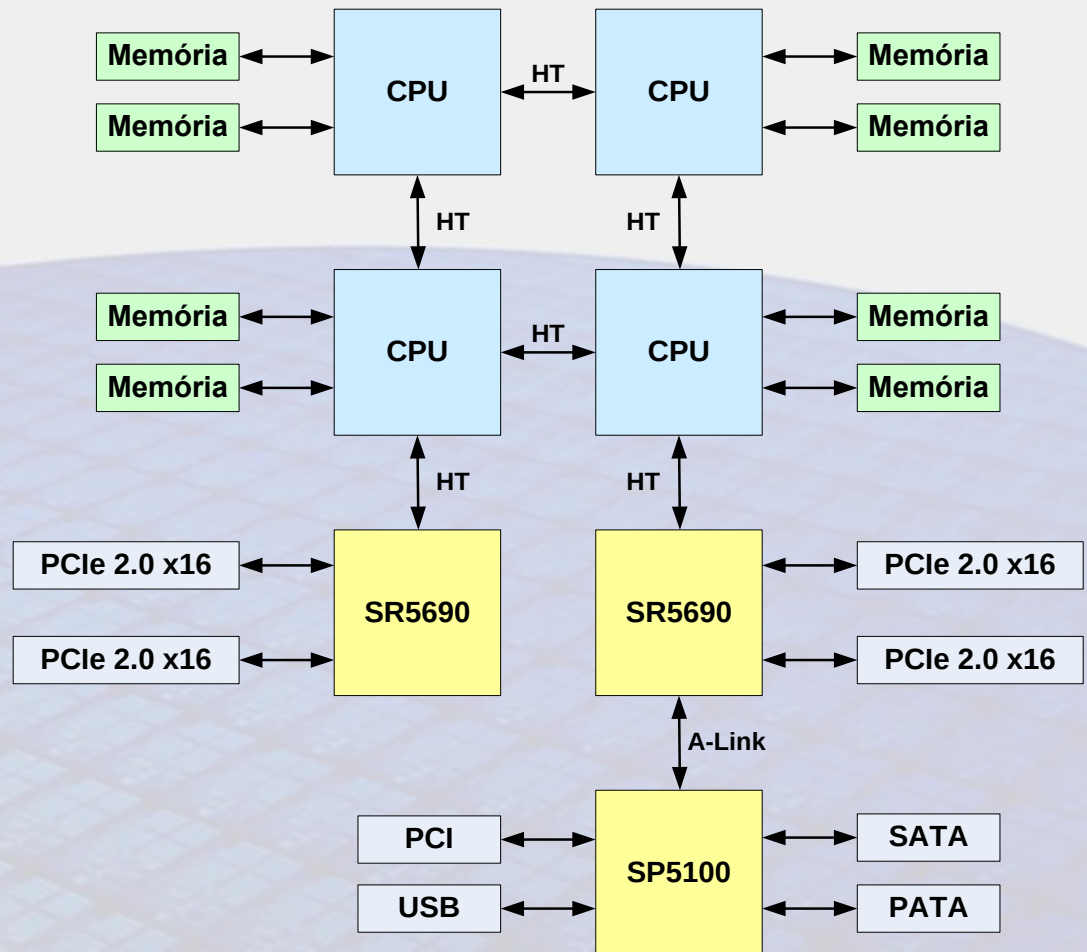
- Memória vezérlő a processzorban

- 2x64 bit DDR2-800
- >32 Gbyte/mag

- Chipszet

- ATI, NVIDIA, Broadcom

- PCIe 2.0 x16



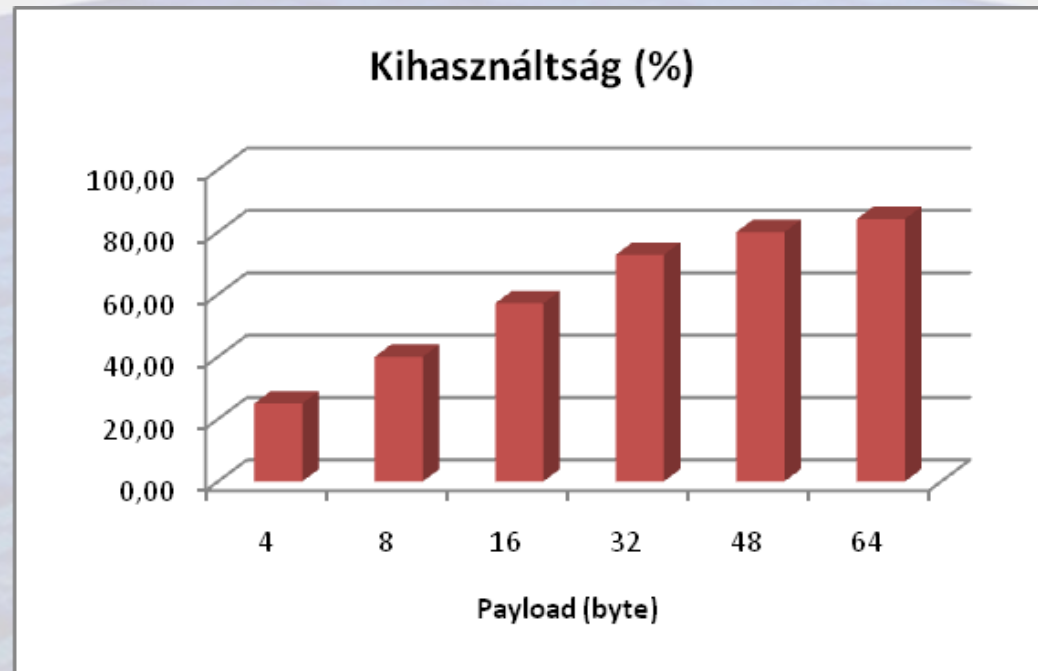
HyperTransport

- „Csomag-orientált” pont-pont kapcsolat
- Egyirányú DDR adatvonalak
 - 2, 4, 8, 16 vagy 32 bit
 - 8 bitenként egy órajel és egy control vonal
 - A két irány szélessége tetszőleges lehet
- Jelenleg 3.1 verzió

	Max. órajel frekvencia	Max. link szélesség	Sávszélesség (16 bit link, egy irányban)
HT 1.1	800 MHz	32	3,2 GByte/s
HT 2.0	1.4 GHz	32	5,6 Gbyte/s
HT 3.0	3.2 GHz	32	12,8 Gbyte/s

HyperTransport

- **HT csomag: 8-12 byte fejléc + 4-64 byte adat**
 - Megegyezik a fizikai rétegben és az átviteli rétegben
 - 8B/10B kódolás opcionális (>HT 3.0)
 - Szorosan csatolt periféria illesztést tesz lehetővé
- **Kihasználtság az átvitt adatméret (payload) függvényében**



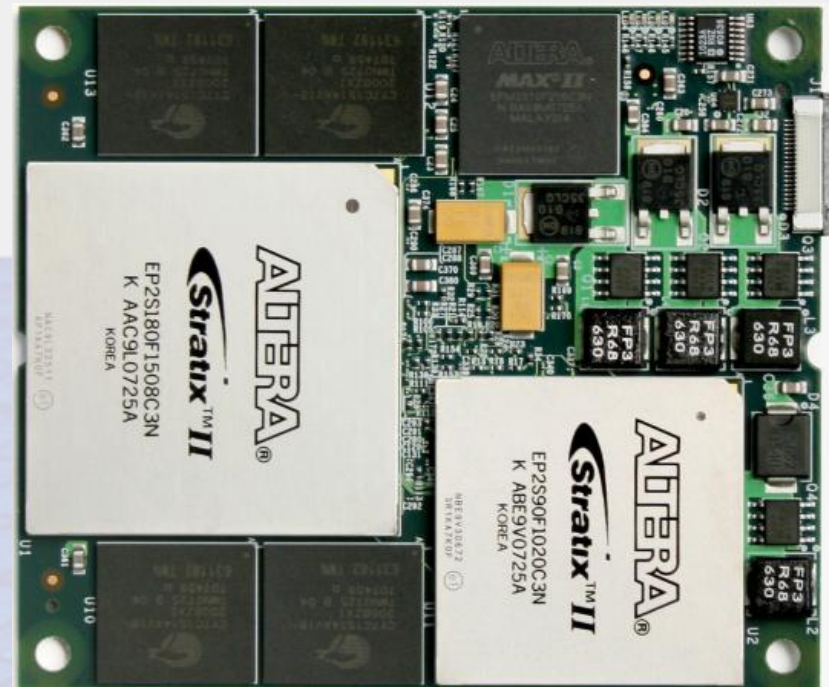
HT csomag (példa)

- **Cmd:** csomag típus
- **SeqID:** csomag csoportosítás, egy csoporton belül a csomagok sorrendje nem változik
- **UnitID:** eszköz azonosító (a másik fél a host bridge)

Bit-Time	7	6	5	4	3	2	1	0
0	SeqID[3:2]		Cmd[5:0]					
1	PassPW	SeqID[1:0]		UnitID[4:0]				
2	<i>Command-Specific</i>							
3	<i>Command-Specific</i>							
4	Addr[15:8]							
5	Addr[23:16]							
6	Addr[31:24]							
7	Addr[39:32]							

XtremeData XD2000F

- Socket-F foglalatba illeszkedő koprocesszor
 - 400 MHz HT link!
- HT interfész FPGA
- Felhasználói FPGA
 - Altera Stratix-II 180
 - 180000 LE, 9 Mbit BRAM, 384 MUL18
- 4 bank QDRII SRAM
 - 36 Mbyte
 - 36 bit, 175 MHz



DRC Accelium

- **Socket-F foglalatba illeszkedő koprocesszor**
 - 400 MHz HT link!
- **FPGA**
 - Xilinx Virtex-5 LX330
 - 330000 LE; 10,6 Mbit BRAM, 128 MUL
- **512 Mbyte RLDRAM**
 - 3,2 Gbyte/s
- **2x2 Gbyte DDR2 DRAM**
 - 6,4 Gbyte/s



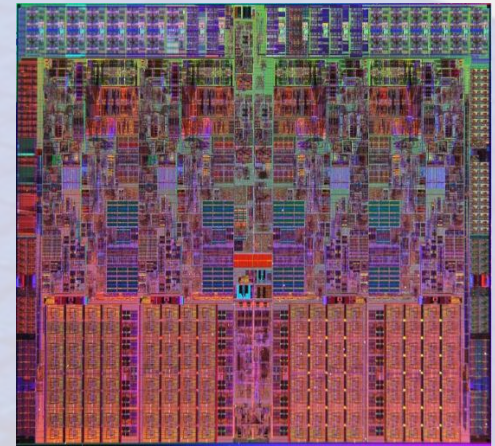
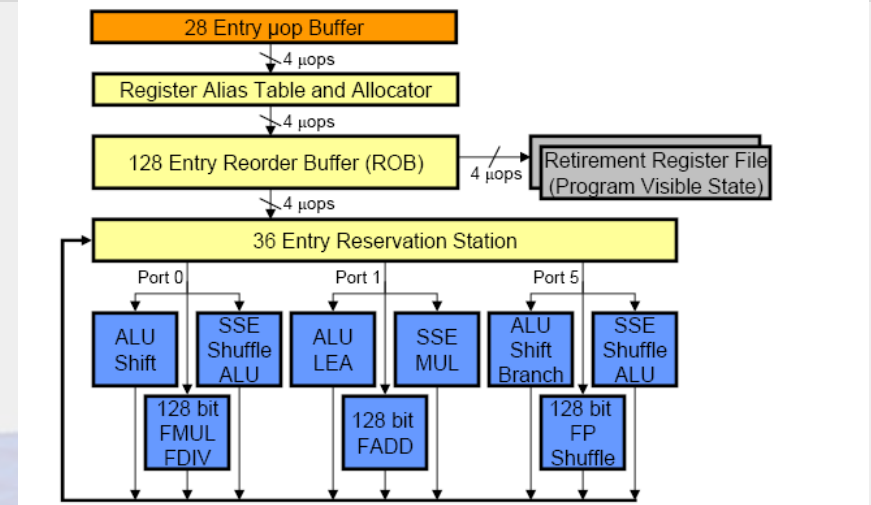
Intel Xeon 3000/5500

■ Intel Nehalem mikroarchitektúrájú

- 16 lépcsős pipeline, SMT
- ALU: 3 végrehajtóegység
- FPU: FMUL/FDIV + FADD
- SSE: 3x128 bit végrehajtóegység
- Memóriavezérlő a processzorban
- DDR3-1333

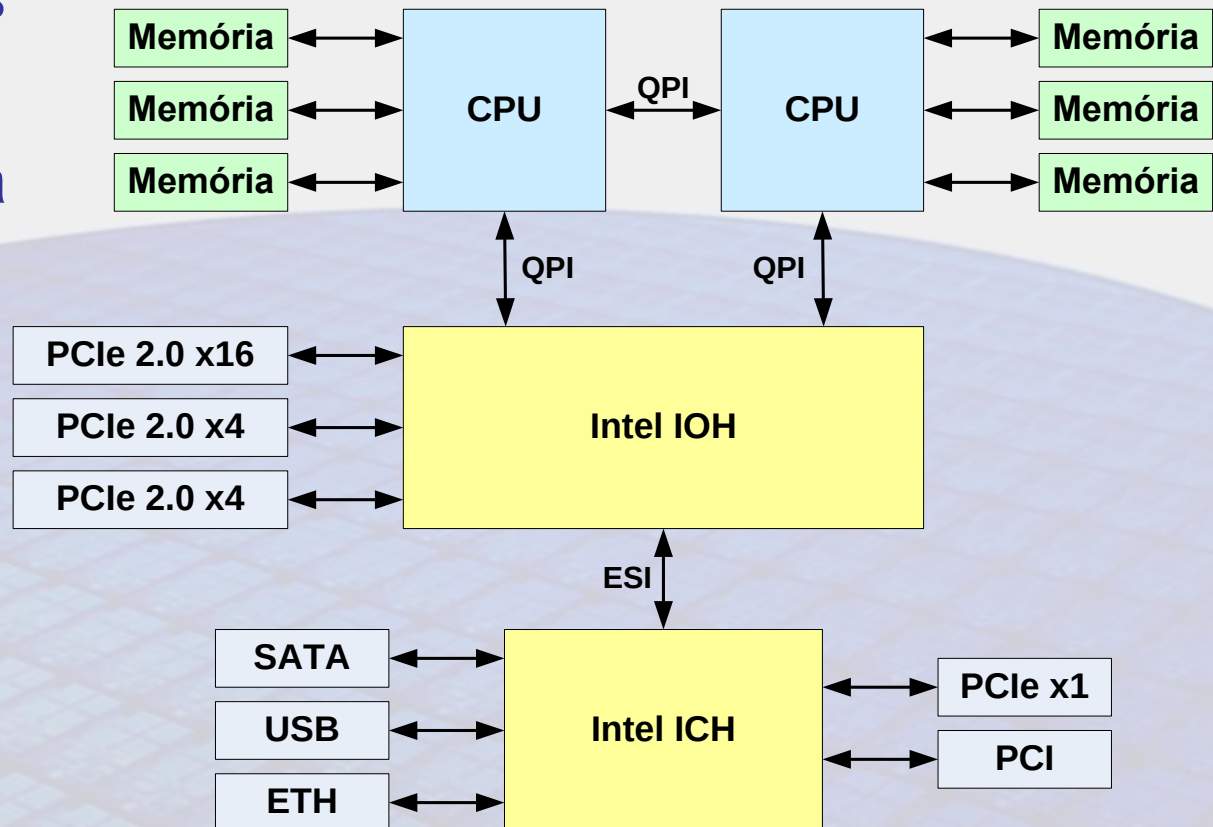
■ Xeon 5500, LGA1366

- Max: 4 mag, 3,33 GHz, \$1600
- 4x64k L1, 4x256k L2, 8 mbyte L3 cache



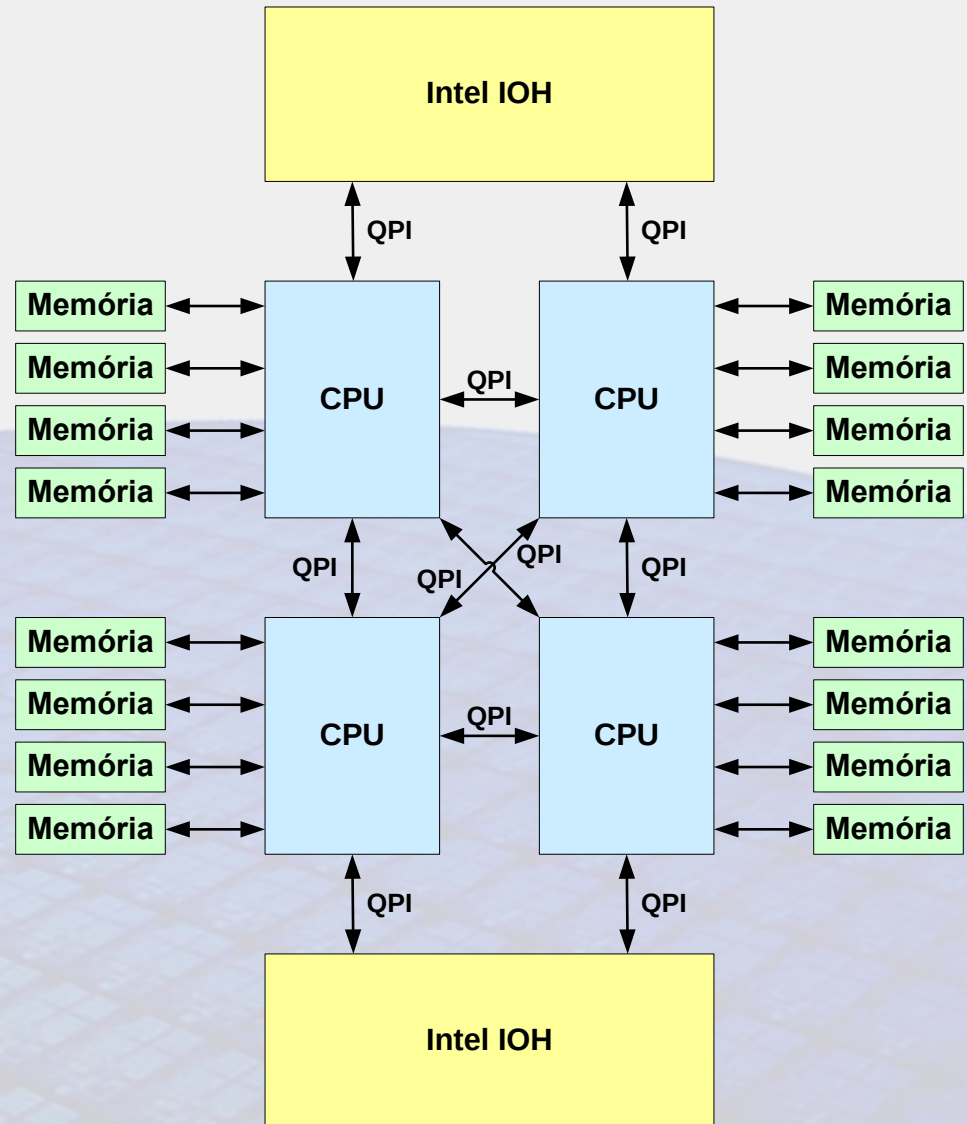
Xeon 5500

- A processzorok QPI linken keresztül kapcsolódnak, gyakorlatilag megegyezik az Opteron struktúrával
 - 19 – 25 Gbyte/s
- 2-utas SMT
- Integrált memória vezérlő
 - DDR3-1333
 - 3x64 bit
 - 72 Gbyte/CPU
- PCIe
 - 2.0, x16



Nehalem-EX

- Több-utas szerverekbe
- 8 mag/processzor
- SMT, 16 szál
- 4 csatornás memória vezérlő
- 4 QPI interfész

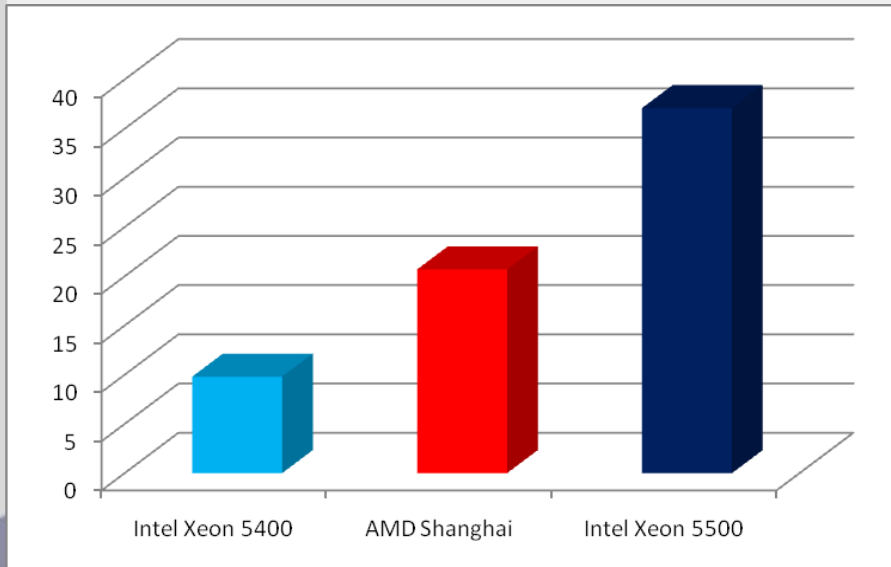


Intel QPI

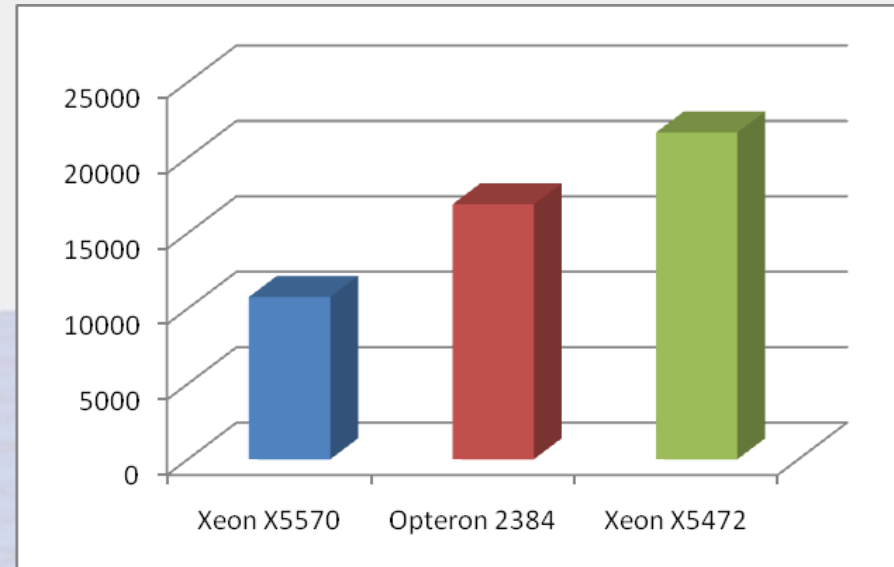
- **Hasonló a HyperTransport-hoz**
 - Egyirányú DDR adatvonalak (jelenleg 20 bit)
 - Külön órajel
 - A két irány adatszélessége eltérő lehet
 - Nincs 8B/10B kódolás
- **Max. 3,2 GHz órajel**
 - 25,6 Gbyte/s sávszélesség
- **Csomag alapú**
 - 8 bit fejléc, 8 bit CRC, 64 bit adat
- **Nallatech QPI gyorsító?**

x86 architektúrák

■ Memória sávszélesség



■ HPC alkalmazás

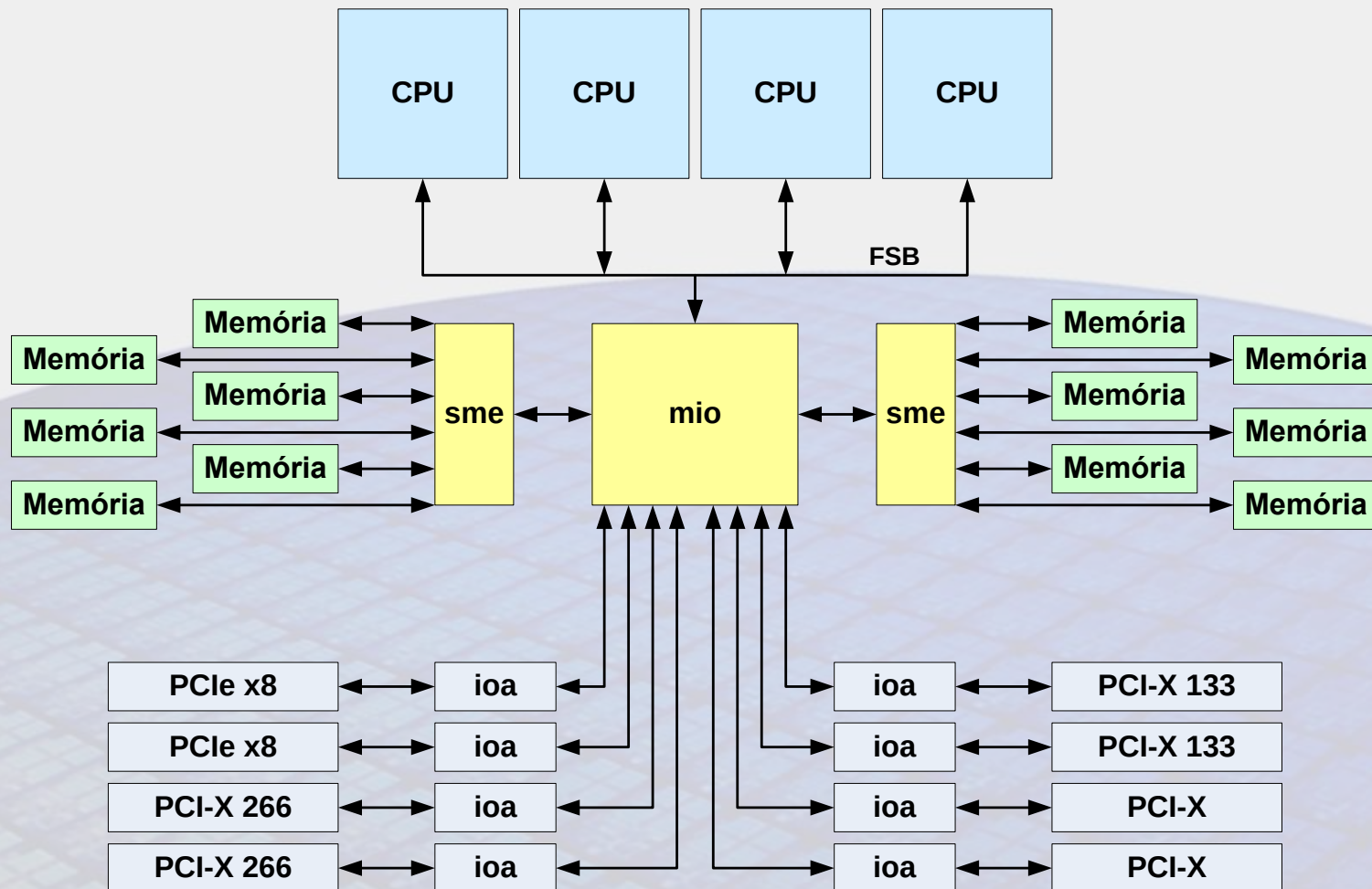


Intel Itanium

- **HP – Intel közös fejlesztés**
- **2001: Merced (~2 éves késés)**
 - ~VLIW utasítás, 2 mag
 - 64 bites architektúra, órajelenként 6 utasítás
 - FSB alapú architektúra, ~1,6 GHz max. órajel
- **2002: Itanium-2**
- **2010: Itanium 9300 (~2 éves késés)**
 - 4 mag, 8 szál (SMT)
 - 2 integrált memória vezérlő + SMI (Scalable Memory IF)
 - Közös infrastruktúra a Nehalem-EX-szel
 - QPI interfész, SMI
 - 1,7 GHz órajel

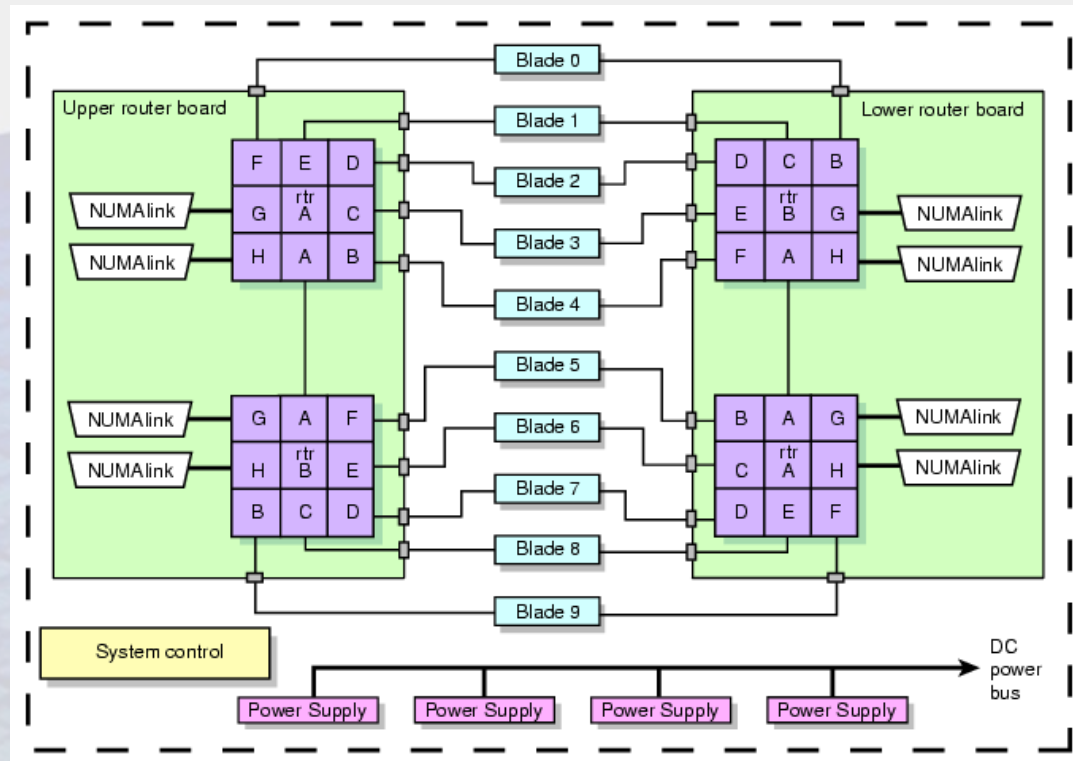
Itanium-2 (HP zx2)

- Chipkészlet: HP zx2



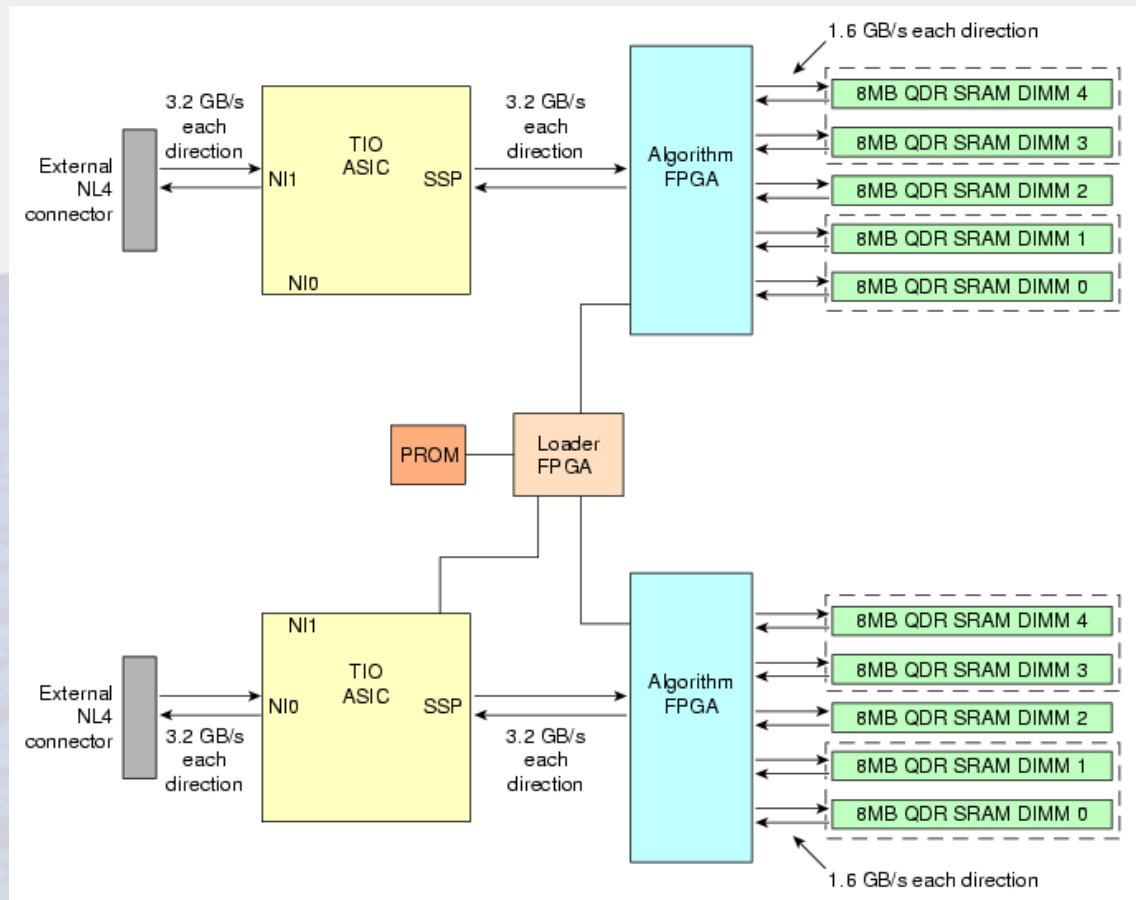
SGI Altix 4700

- Itanium-2 alapú szerver család
- SGI NUMalink-4
 - Elosztott osztott memória hozzáférés, kis késleltetés, jó skálázódás



SGI RC100

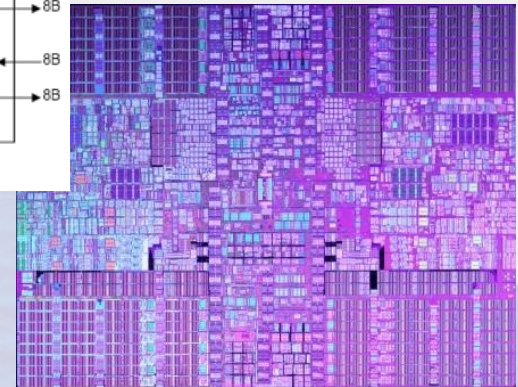
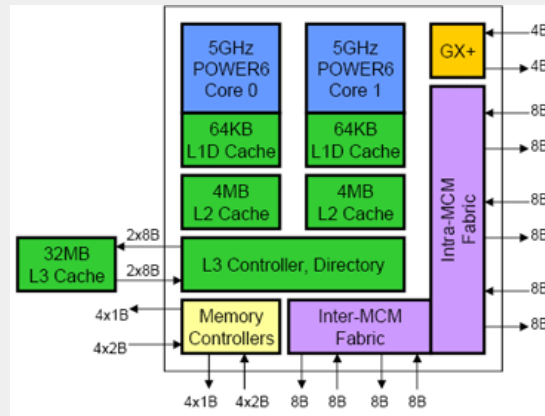
- Altix 4700 rendszerbe illeszkedő, FPGA alapú gyorsító
- 2x Virtex-4 LX200
- 2x 5 csatorna QDR SRAM
- NUMalink-4 kapcsolat TIO IC-n keresztül



IBM POWER6

■ POWER6 architektúra

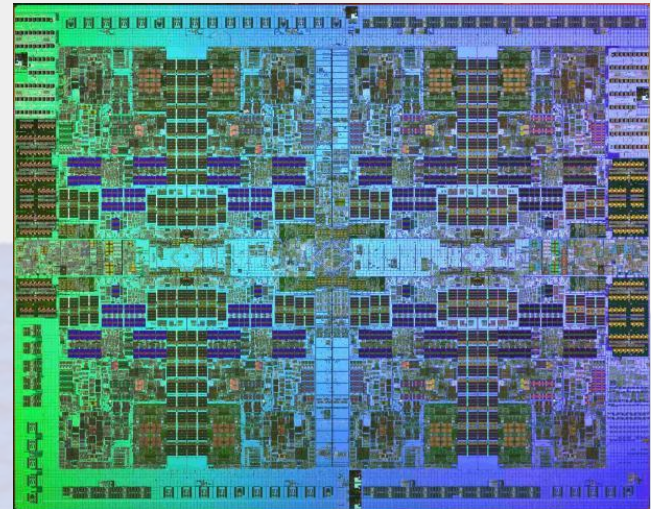
- 2 CPU mag
- In-order, 2x SMT
- 2x ALU
- 2x FPU
- Altivec vektor egység
- 5 GHz órajel
- 2x64k L1, 2x4M L2, 32M L3 (MCM) cache
- On-chip memória vezérlő (DDR2/DDR3)



IBM POWER7

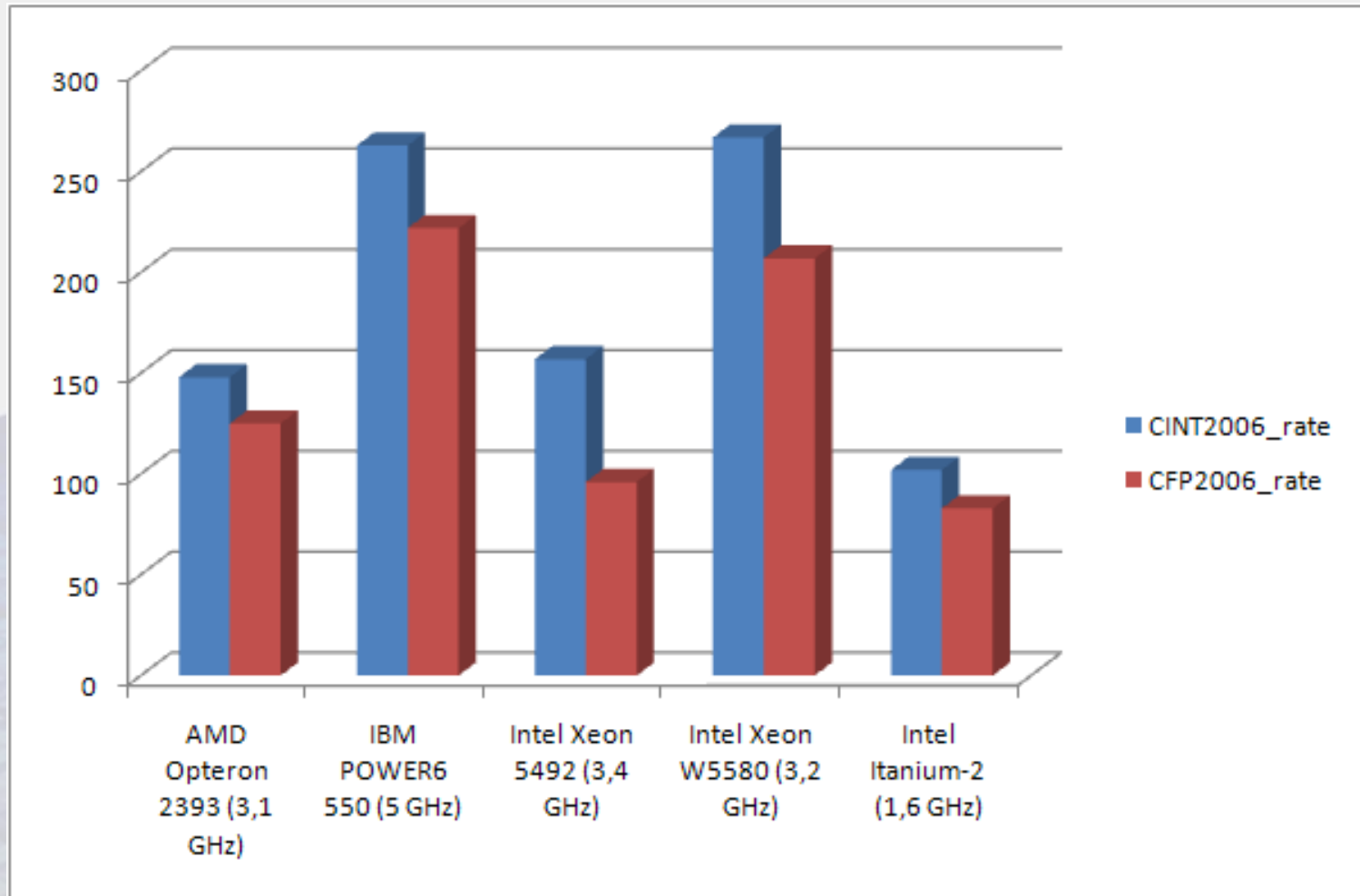
■ POWER7 architektúra

- Out-of-order végrehajtás
- 4/6/8 CPU mag
- 3 – 4 GHz órajel
- 4x SMT magonként
- 2 ALU
- 4 FPU
- 1 AltiVec
- Magonként 32k L1, 256k L2 osztott 32M L3
- 2x DDR3 memória vezérlő
- Max. 32 utas kialakítás



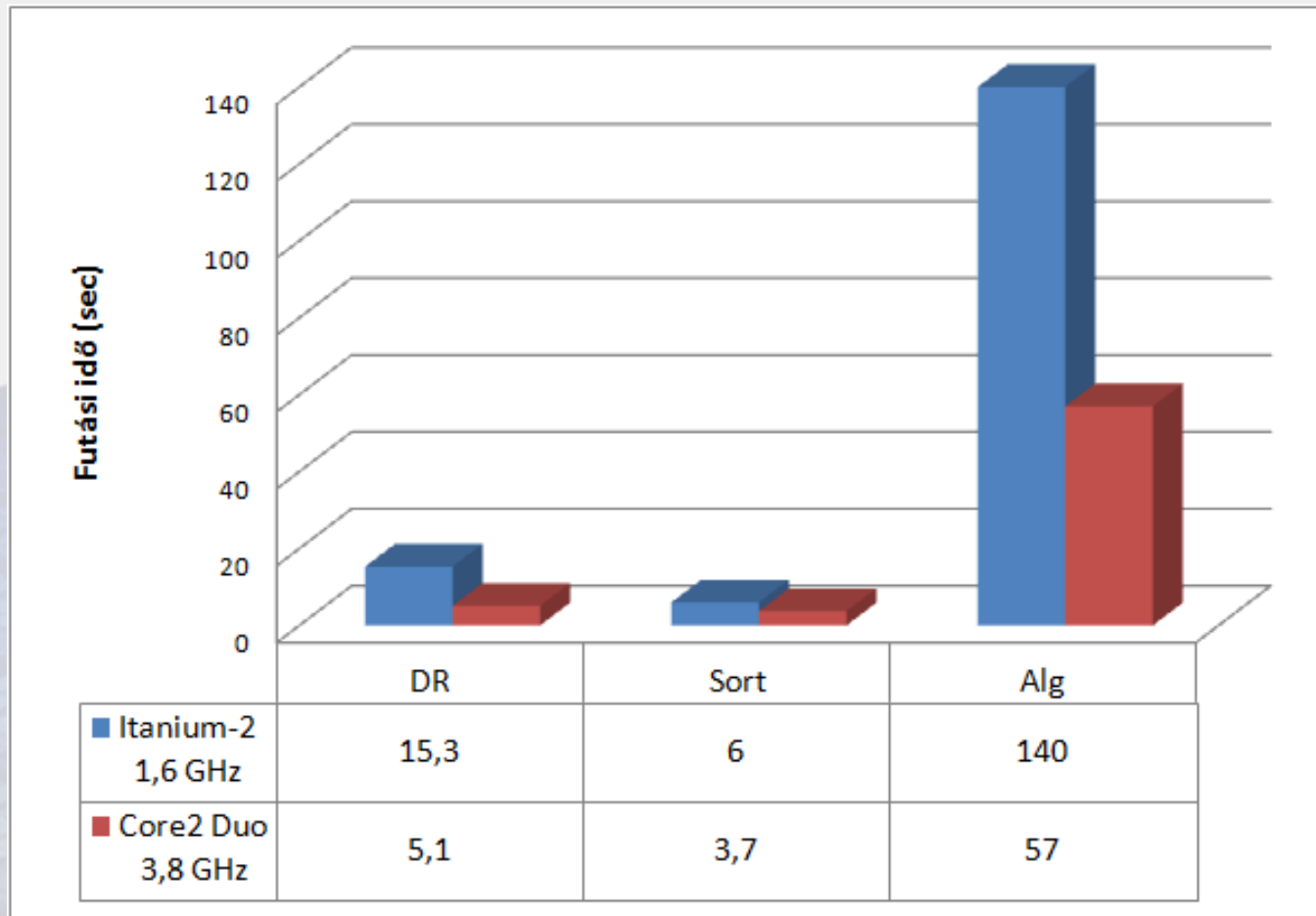
SPECint/SPECfp

- SPECint 2006 & SPECfp 2006 eredmények (8 mag)



Itanium – Core2

- Bioinformatikai algoritmus



PCI

- **PCI, PCI-X: „igazi” busz struktúra**
 - Multiplexált cím/adatvonal
 - Half-duplex
 - Multi-master, DMA
 - 4 interrupt vonal

Szabvány	Adatszélesség	Frekvencia	Sávszélesség Mbyte/s
PCI	32/64 bit	33 MHz	133/266
PCI 2.1	32/64 bit	66 MHz	266/528
PCI-X 1.0	64 bit	66/133 MHz	528/1056
PCI-X 2.0	64 bit	266/533 MHz	2112/4224

PCI

- **Fontosabb jelek**

- AD: cím/adatvonal
- C/BE#: parancs vagy byte engedélyezés
- FRAME#: aktív átvitel
- IRDY#: initiator (master) ready
- TRDY#: target ready

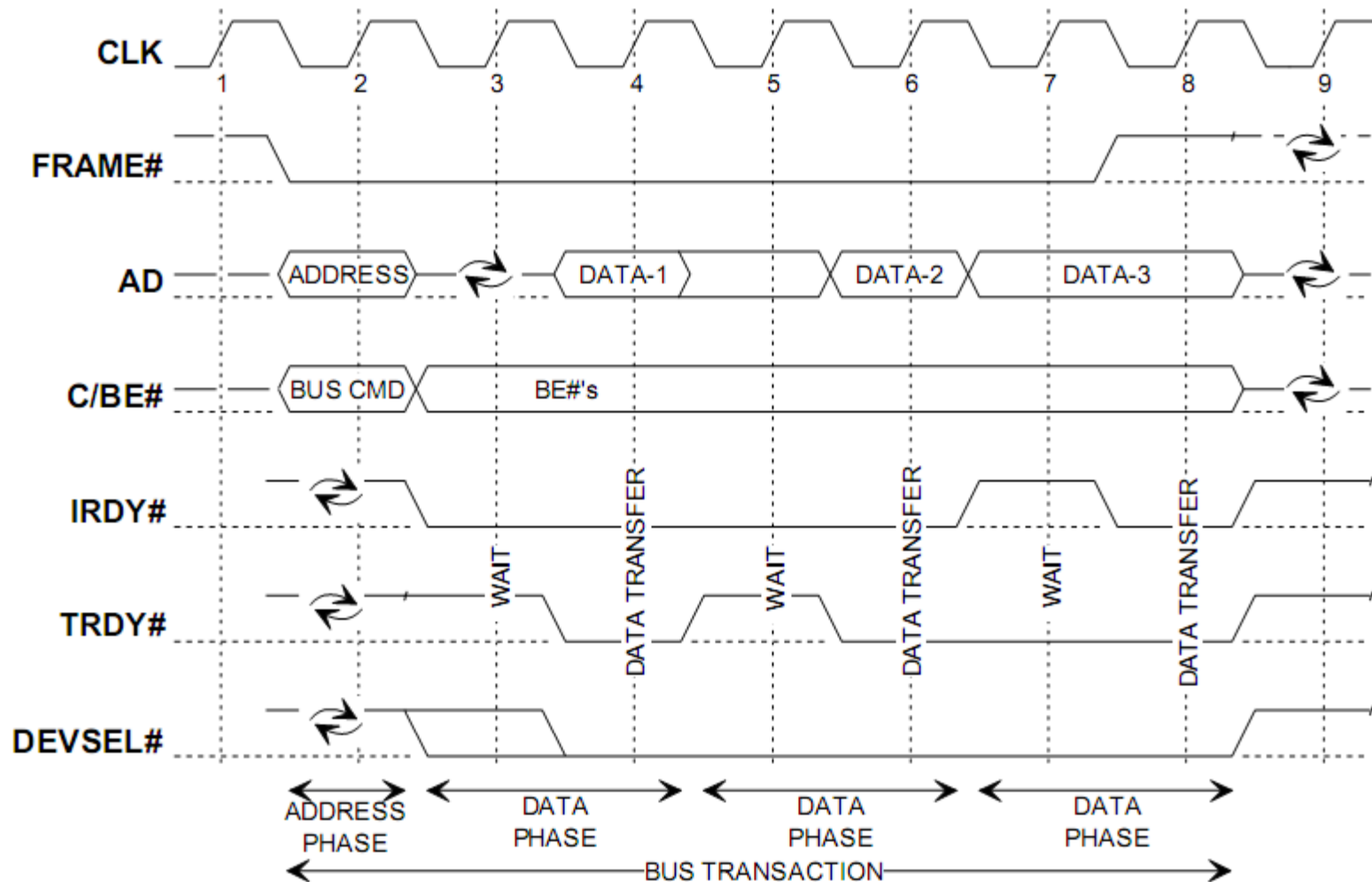
- **Perifériák**

- Target: nem kezdeményezhet buszciklust
- Bus master: lehet master a buszon
- A központi PCI vezérlő tipikusan NEM tartalmaz DMA vezérlőt!

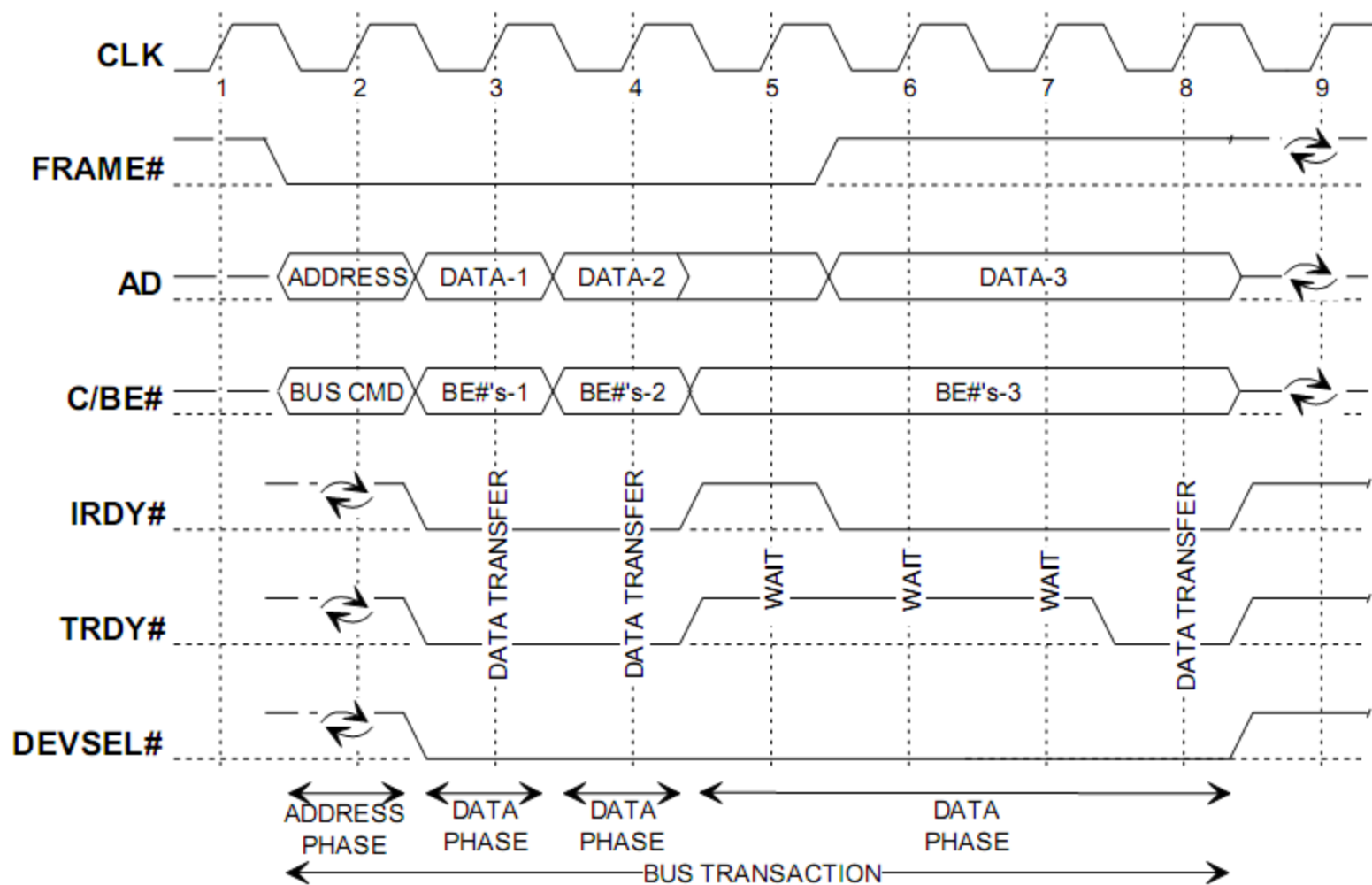
PCI

- **Parancsok (BUS CMD)**
 - Interrupt Ack
 - Special Cycle
 - I/O Read és Write
 - Memory Read és Write
 - Configuration Read és Write
 - Dual Address Cycle (pl.: 64bit címzés)

PCI olvasási ciklus



PCI írási ciklus



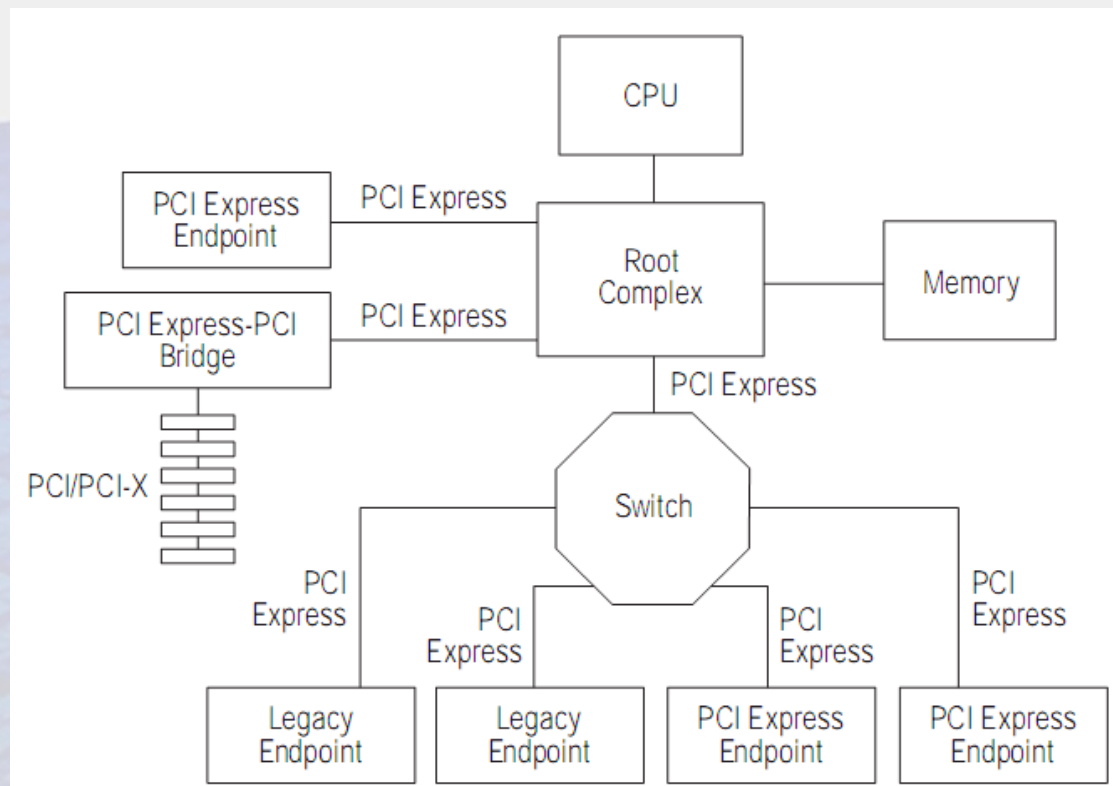
PCI Express

- Soros, pont – pont összeköttetés; helyreállított órajel
- Csomag alapú átvitel
- Szoftveresen kompatibilis a PCI-jal
- 1, 2, 4, 8 vagy 16 lane
 - PCIe 1.1: 2,5 Gbit/s (8b/10b)
 - PCIe 2.0: 5 Gbit/s (8b/10b)
 - PCIe 3.0: 8 Gbit/s (128b/130b)
- **Kommunikáció**
 - Requester
 - Completer

	PCIe 1.0 (Gbyte/s)	PCIe 2.0 (Gbyte/s)
x1	0,25	0,5
x4	1	2
x8	2	4
x16	4	8

PCIe architektúra

- **Root Complex: központi vezérlő**
 - Kapcsolat a CPU-val és a memóriával
- **Endpoint: periféria**
 - Completer
 - Completer + Requester
- **Nincs központi DMA vezérlő**

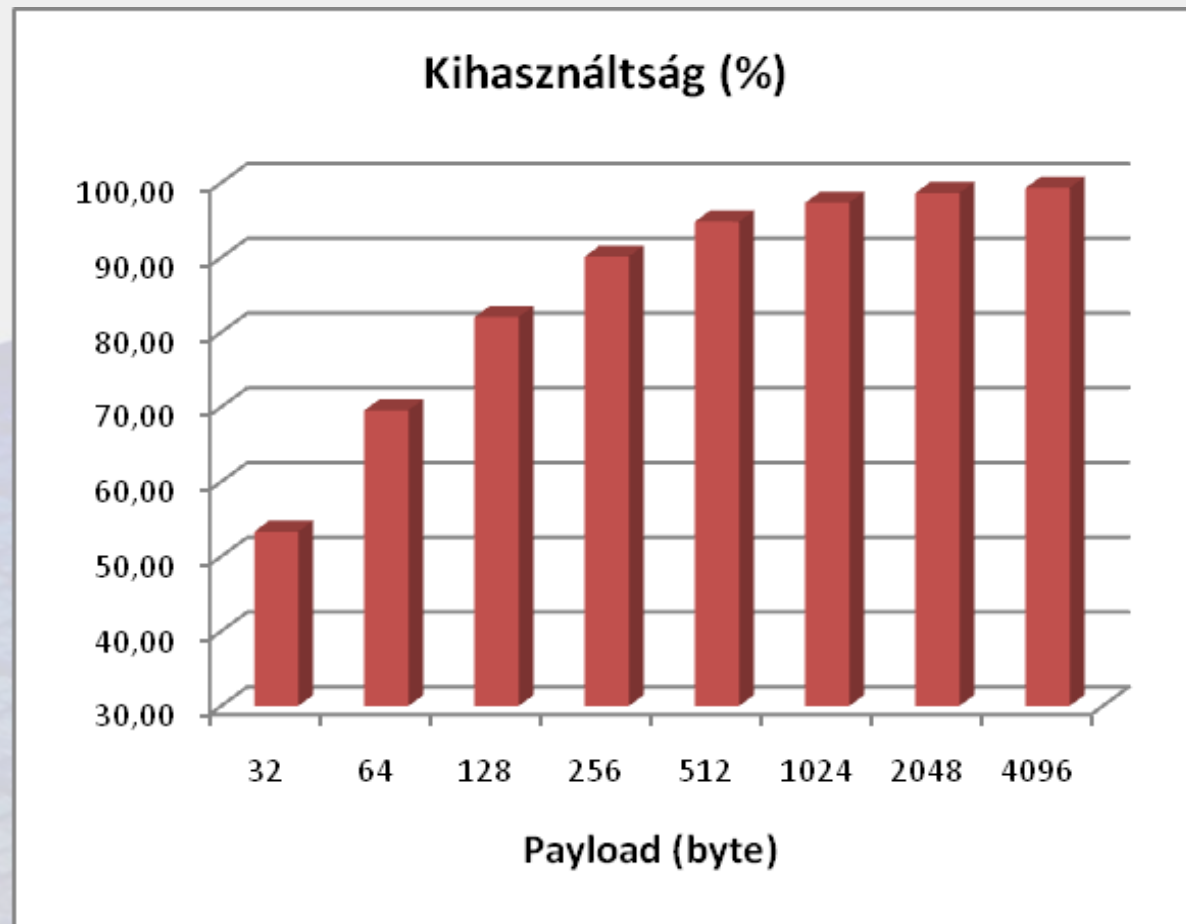


PCle

- **Fizikai réteg: differenciális érpár**
 - Framing + CRC (2x64 bit)
- **Átviteli réteg**
 - Fejléc: 3-4 DW
 - Opcionális CRC: 1 DW
 - Data payload: 128/256/512/1024/2048/4096 byte
 - Root Complex & Endpoint támogatás kell!

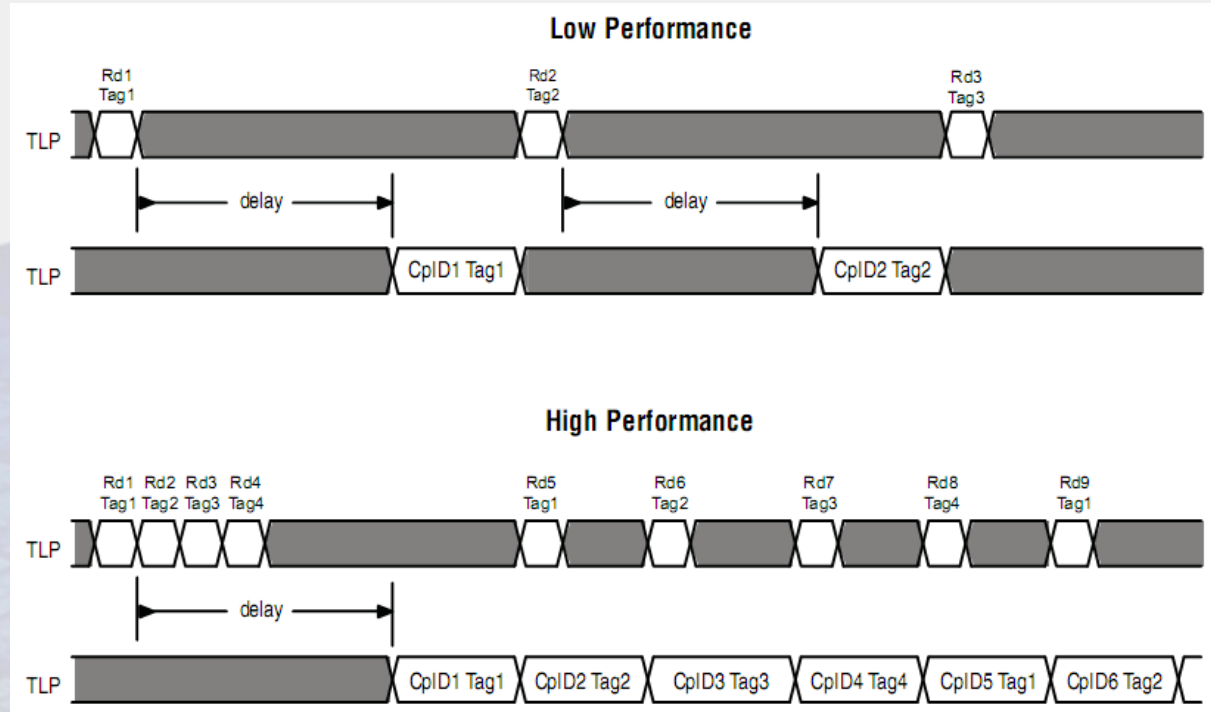
PCle

- Egy csomagban legfeljebb Data Payload-nyi adat
 - Nagyobb méretű request-ek több csomagban



PCIe

- A PCIe késleltetése viszonylag nagy
 - De újabb Request-ek elküldhetők a Completion megérkezése előtt

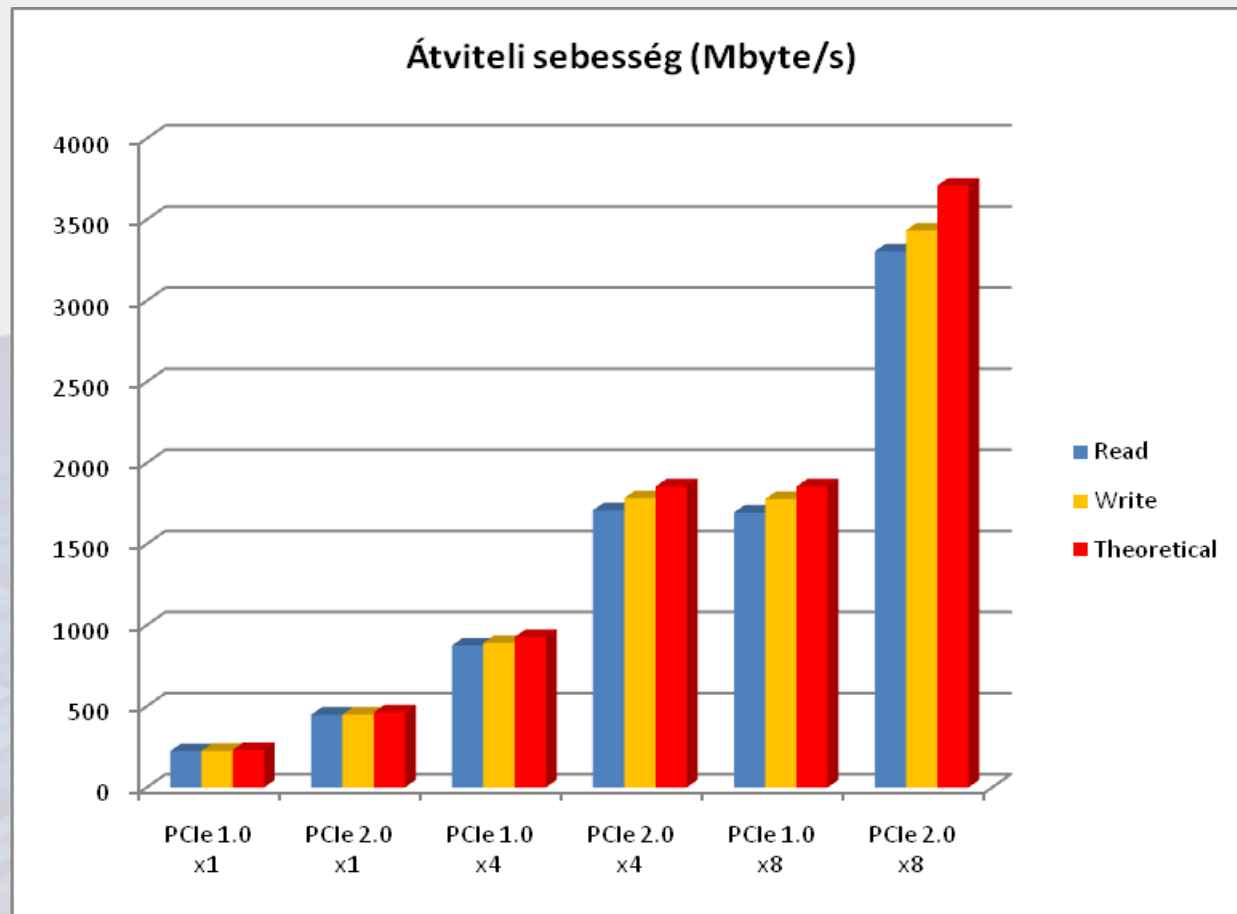


PCle

- **Egy Request-re adott Completion**
 - Egy csomagban
 - Több csomagban – a csomagok sorrendje címsorrendben
- **Több, címfolytonos Request**
 - A Completion csomagok sorrendje NEM feltétlenül felel meg a címsorrendnek

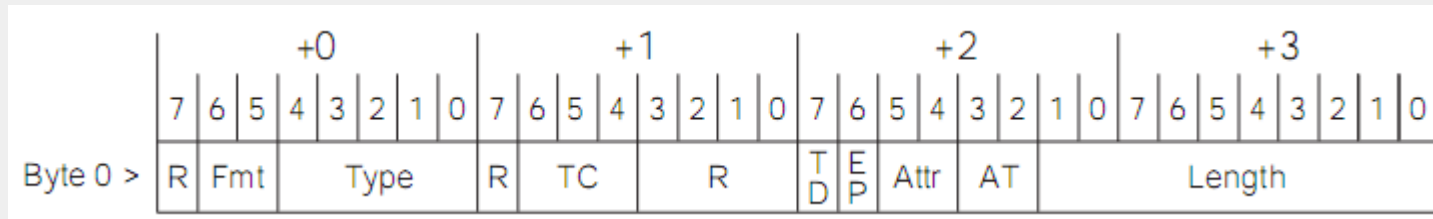
Valós átviteli sebesség

- PCle 1.1 és 2.0 buszon elérhető sebesség (Intel X58 chip, 256 byte Payload)



PCIe TLP

■ Fejléc 1. DW (minden csomagban)

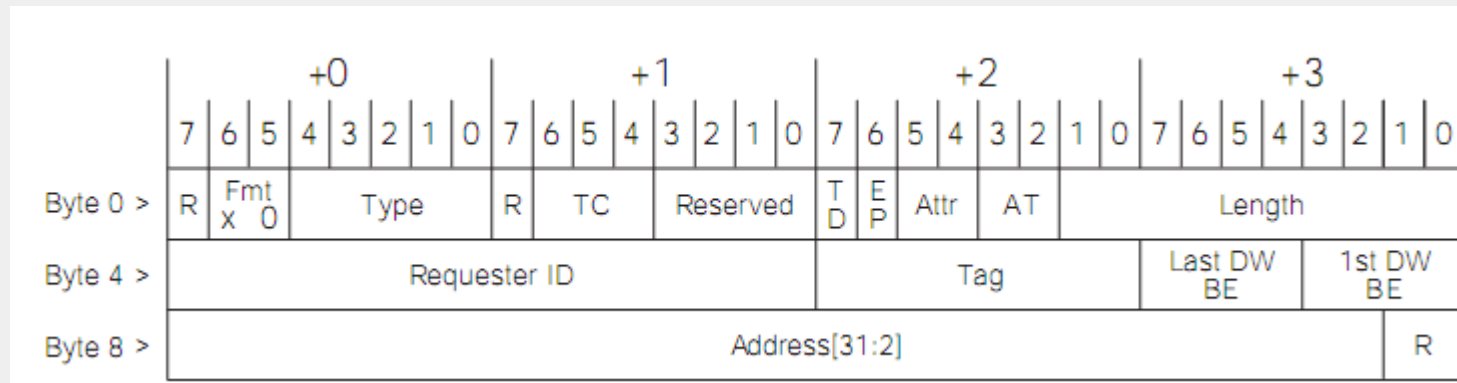


Fmt	Type	
00 01	0_0000	Memory Read Request
10 11	0_0000	Memory Write Request
00	0_1010	Completion, nincs adat
10	0_1010	Completion, adattal

- Length: 1...1024 DW (< Data Payload)

PCIe – 32 bites Request

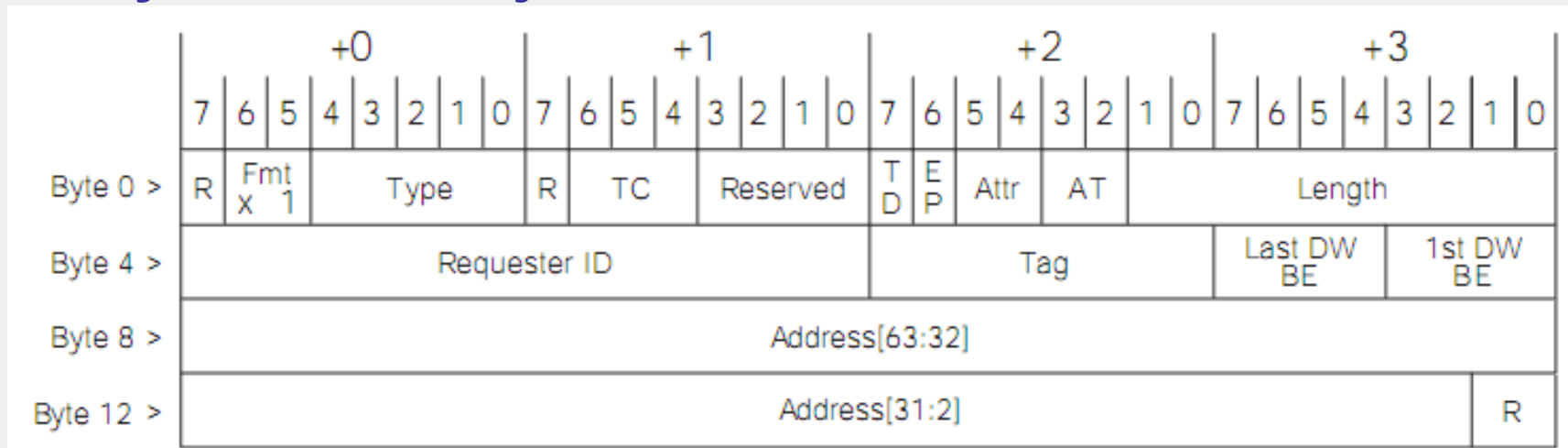
■ TLP fejléc



- Requester ID: Bus number + Device Number + Function number
- Tag: Minden „kintlévő”, Completion-t igénylő üzenet egyedi azonosítója (kintlévő üzenetek száma < 32)
- BE: byte engedélyezés az első és utolsó DW-re
- Address: cím (DW cím)

PCIe – 64 bites Request

- 3 helyett 4 DW a fejlécben



PCIe - Completion

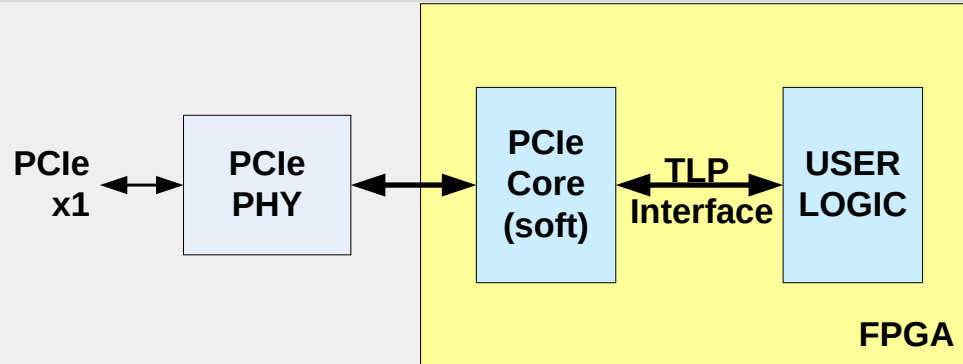
Completion TLP fejléc

	+0								+1								+2								+3								
	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0	
Byte 0 >	R	Fmt x 0		Type				R	TC		Reserved						T D	E P	Attr		AT 0 0		Length										
Byte 4 >	Completer ID																Compl. Status		B C M		Byte Count												
Byte 8 >	Requester ID																Tag						R	Lower Address									

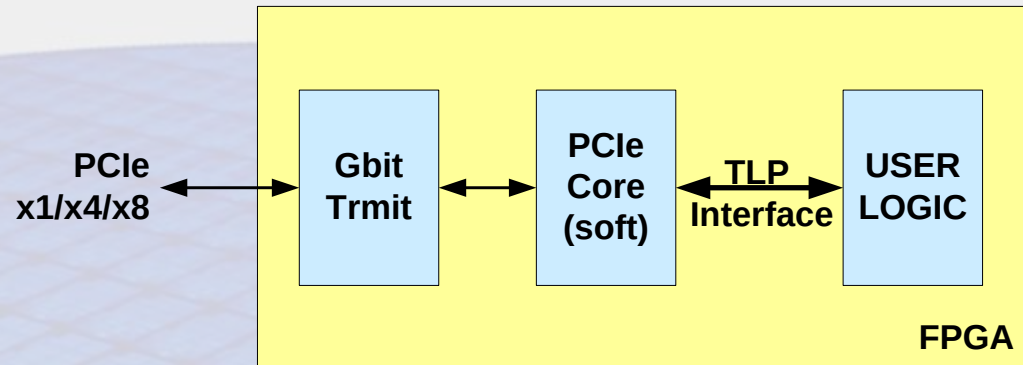
- **Tag:** A Request-tel megyező Tag
- **Byte Count:** A Request-ből még hátralévő byte-ok száma

PCIe FPGA-val

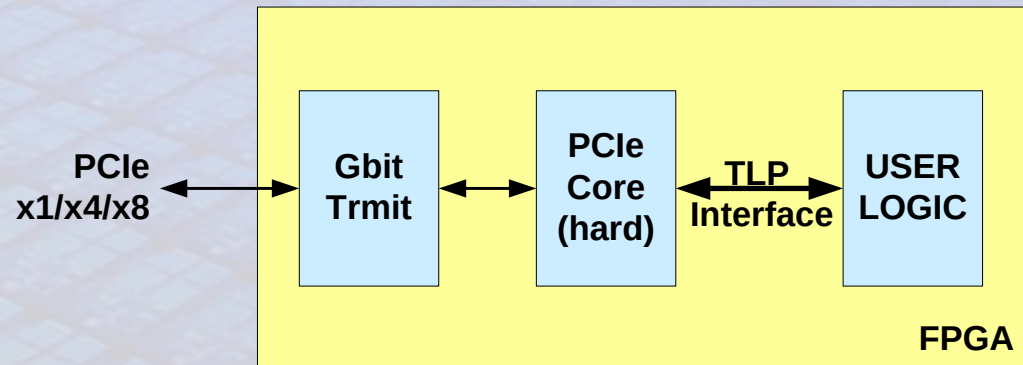
- Külső PHY
Soft-core MAC
(minden FPGA)



- Belső Gbit transmitter
Soft-core MAC
(pl. V2P, V4)



- Belső Gbit transmitter
Hard-core MAC
(pl. SP6, V5, V6)



FPGA megvalósítások

FPGA	x1	x4	x8	PCIe 2.0	hard IP
Altera Arria II GX	X	X	X		X
Altera Cyclone IV GX	X	X			X
Altera Startix IV GX	X	X	X	X	X
Altera Arria GX	X	X			
Altera Stratix II GX	X	X	X		
Xilinx Spartan-6	X				X
Xilinx Virtex-II Pro	X	X	X		
Xilinx Virtex-4 FX	X	X	X		
Xilinx Virtex-5	X	X	X		X
Xilinx Virtex-5 FXT	X	X	X	X	X
Xilinx Virtex-6	X	X	X	X	X
Lattice ECP2M	X	X			
Lattice ECP3	X	X			
Lattice SCP	X	X			X

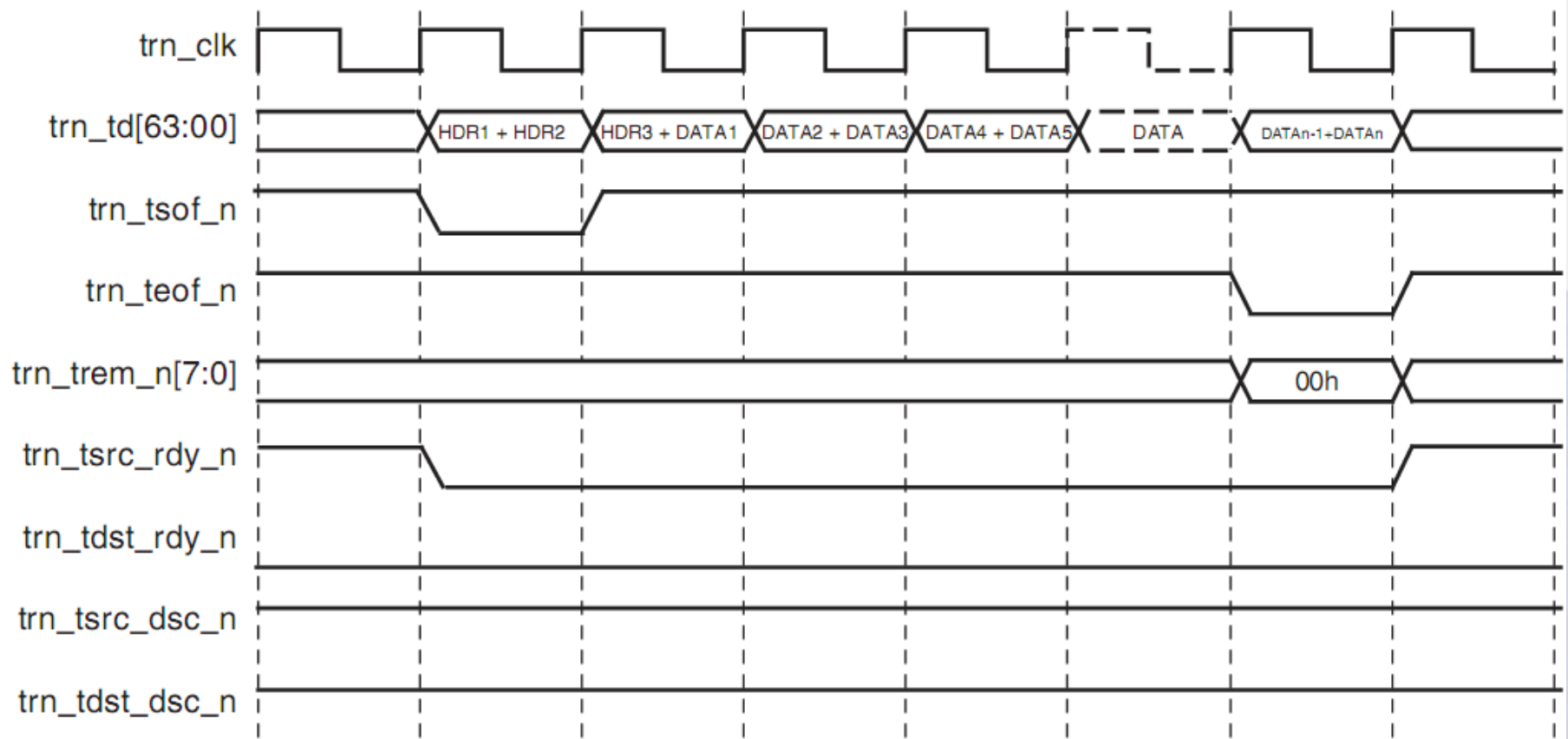
PCIe FPGA interfész

- Gyakorlatilag Memory Read, Memory Write és Completion csomagok fogadása és generálása szükséges
- TLP interfész: 64 bites FIFO IF

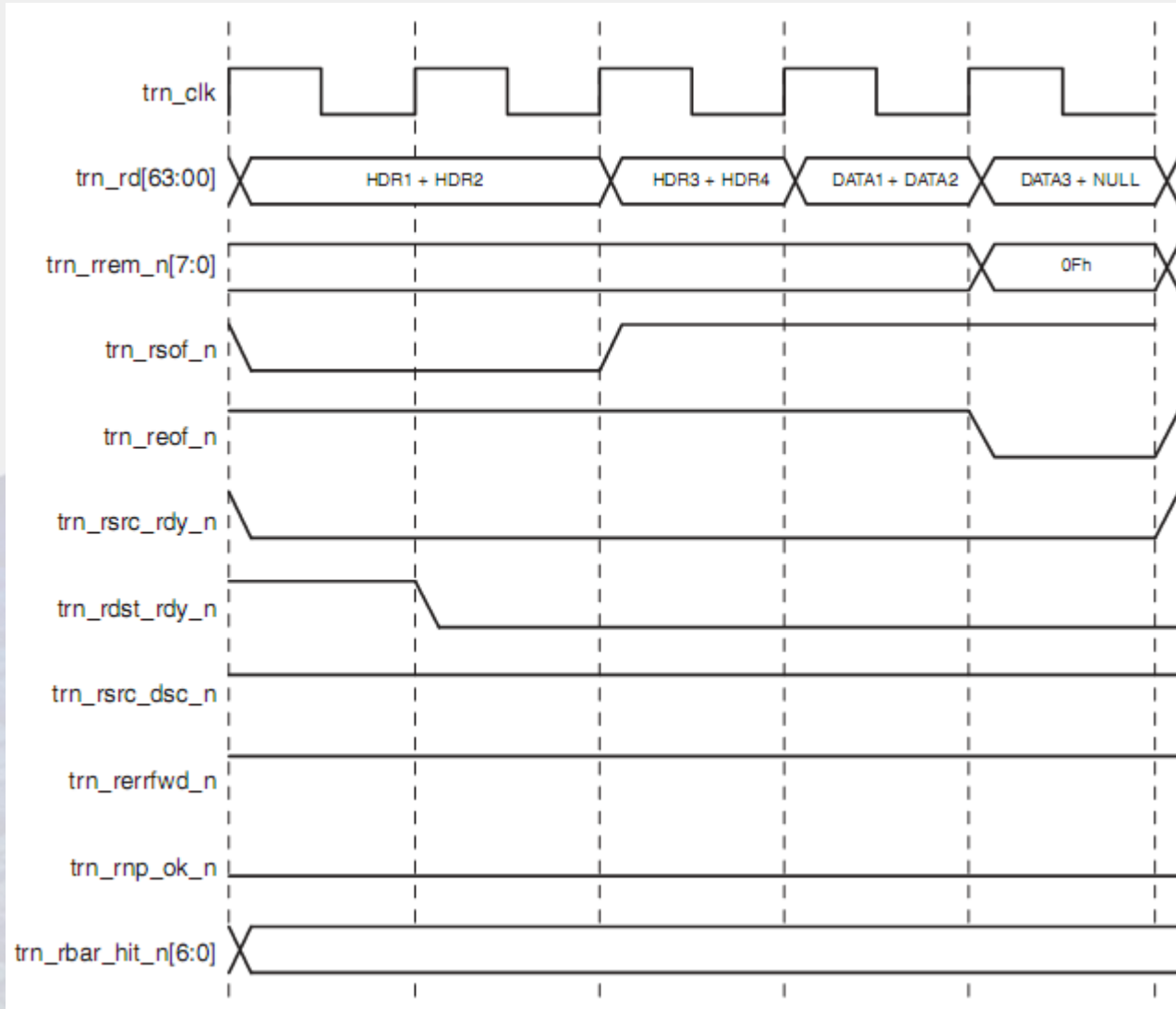
	+0 [63:56]		+1 [55:48]		+2 [47:40]		+3 [39:32]		+4 [31:24]		+5 [23:16]		+6 [15:8]		+7 [7:0]			
Byte 0 >	R	Fmt x 0	Type	R	TC	Rsvd	T D	E P	Attr	R	Length		Requester ID		Tag	Last DW BE	1st DW BE	
Byte 8 >	Address[31:2]										R	Data 0						
Byte 16 >	Data 1										Data 2							
Byte 24 >	TLP Digest										(no data)							

- 1/4/8 lane: 62,5/125/250 MHz

PCle Virtex5 - Transmit

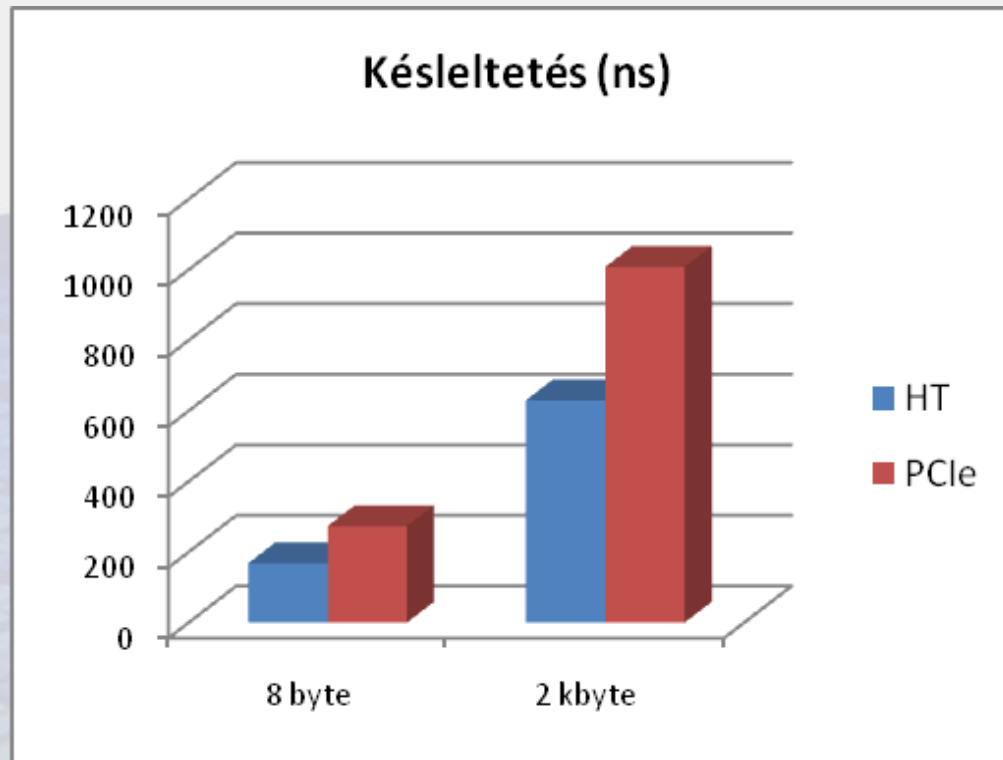


PCIe Virtex5 - Receive



HT vs. PCIe - késleltetés

- HT: Opteron „natív” interfész (szorosan csatolt)
- HT: kevesebb járulékos információ ugyanannyi adathoz
- PCIe 8b/10b kódolás késleltetése

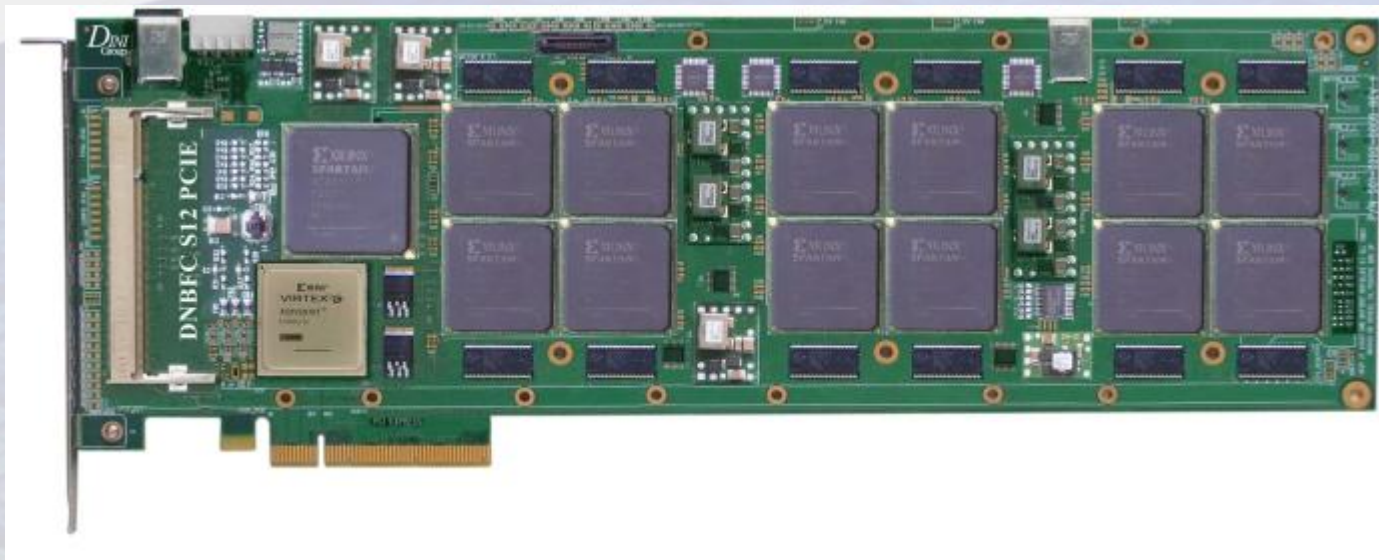


Elérhető sávszélesség



DINI Group DNBFC_S12_PCl_e

- HPC alkalmazásokhoz fejlesztett „olcsó” megoldás
 - 12 db Spartan-6 LX150 FGPA (~90000 LUT/FPGA)
 - 256 Mbyte DDR3 SDRAM minden FPGA-hoz
 - Virtex-6 FPGA a PCIe kapcsolathoz
 - Az FPGA-k között 64 bites buszok



BME MIT ART1

- **Xilinx Virtex-5 SX240 FPGA**
 - 150000 LUT, 18Mbit BRAM, 1056 DSP egység
 - 8-lane PCIe 1.1 csatlakozó
- 4 db, független DDR2 SODIMM foglalat
- DVI, Ethernet, SATA csatlakozó

