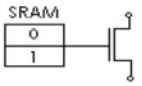
ZH 2

2021. december 8., szerda 15:33

Specify the advantages and disadventages of SRAM based FPGA physical programming method. (2p)

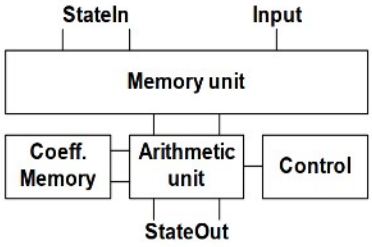
a.) SRAM cell

- Properties
  - Reprogramable infinite times (static RAM)
  - Switching off the content of the SRAMs will be lost
  - Switching on the program has to be downloaded
  - SRAM cell is connected to the gate of a pass transistor. The transistor can be in open or closed state. Interconnections and MUXsare also stored .
  - 1 bit storing in SRAM (min. 6 tranzistors)
  - many tranzistors (standard CMOS), large size, large dissipated power
  - SRAMs do not need refreshing
  - Large 0.5-2 kΩ resistance
  - 10-20 femtoF large parasitic capacitance



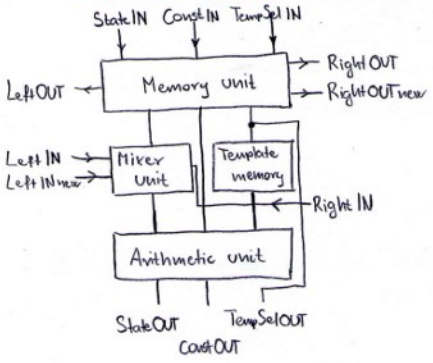
2. Give the FALCON FPGA processor architecture, list the paralellisms in the operation! (3p)

Optimized Falcon-ML emulated digital CNN-UM



- Multi-layer structure (10+ layers)
  - General multi-layer structure is too large
- Optimized application specific architecture
- Memory unit for local storage of state values

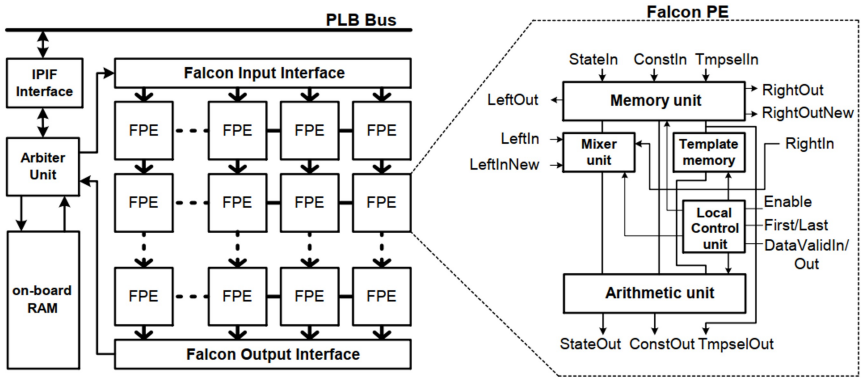
A Falcon processing unit



Configurable parameters
- independent state, template, constant width
- number of templates
- Size of the templates
- Width of the cell array slice
- Number of layers
- Number of processor cores and their arrangement

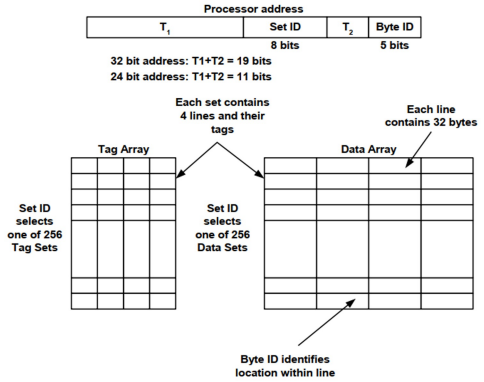
Parallelisms: multiple Falcon blocks can be synthetized on FPGA, arithmetic operations can be parallelized.

- **FPE:** kibővített FALCON Processzáló Elem



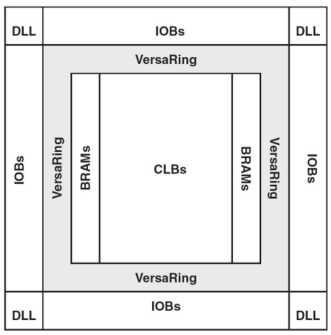
3. Draw the block diagram of a 4-way set associative 32Kbyte cache with 64byte cache line size! How many bits from a 64bit address should be stored in the cache? (5p)

Organization of a 4-way set associative cache



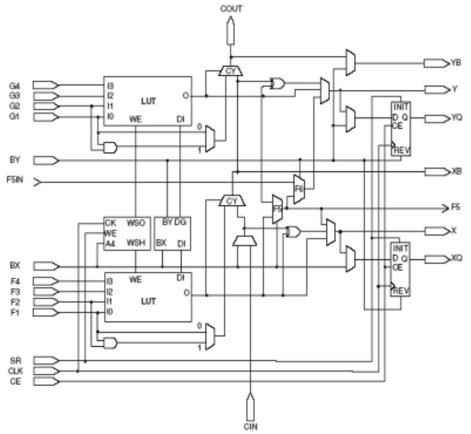
4. Draw the architecture of a Xilinx Virtex FPGA Slice! (3p)

b.) Virtex architecture



- CLB: configurable logic blocks
- BRAM: Block SelectRAM
  - dual-portos 4096-bit RAM
- DLL: Delay-Locked Loop
  - Clock management
  - Clock dividing/multiplying 1.5, 2, 2.5, 3, 4, 5, 8, or 16
  - Phase sift: 0º, 90º, 180º, 270º
- IOB: Input/Output Block
  - 13 different I/O standards are supported (for example: LVDS)

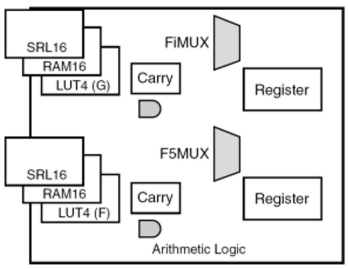
Virtex: Slice



- Look-Up Table
  - 4-input LUT
  - 16 x 1-bit sincronous RAM
  - 16-bit shift register
- Storing elements
  - Edge triggerd D flip-flops
- Additional components
  - F5 multiplexer
    - 4:1 multiplexer
  - F6 multiplexer
    - 8:1 multiplexer
- Aritmetic logic units
  - Dedicated carry logic
  - Dedicated AND gates

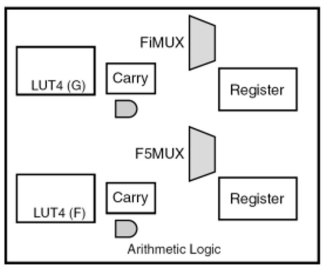
20

Resources in a slice



SLICEM

M: Slice for Memory (Logic, Distributed RAM, Shift Register)



SLICEL

DS312-2_13_020905

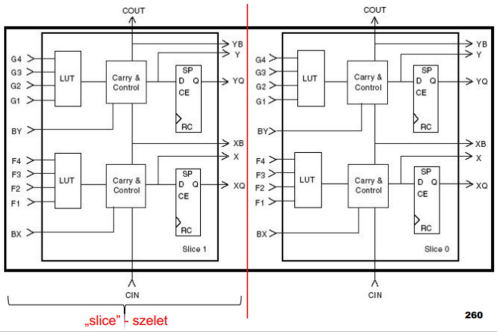
L: Slice only for Logic

21

Slice M:
- LUT:
  - → 4 input LUT
  - → 16x1 bit synchronous RAM
  - → 16 bit shift register
- Storing elements
- Multiplexers
- ALU-s
  - → dedicated carry logic
  - → dedicated AND gates

Slice L:
- LUT:
  - → 4 input LUT
- Storing elements
- Multiplexers
- ALU-s
  - → dedicated carry logic
  - → dedicated AND gates

Virtex (CLB) Configurable Logic Block

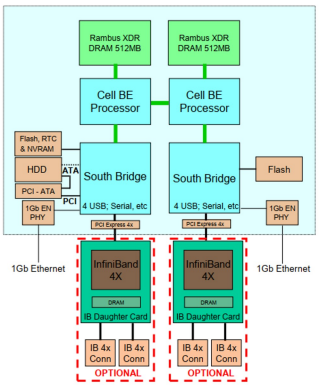


260

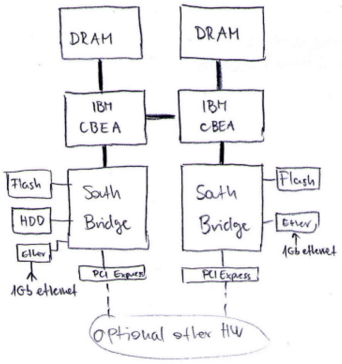
5. Draw the blockdigram of the IBM Cell architecture and specify the main parameters of the architecture!(3p)

CELL Blade

- Cell BE Processor Blade **(~500GFLOPS peak)**
  - Dual 3.2GHz Cell BE Processor Configuration
  - 1GB XDRAM (512MB per processor)
  - Blade-mounted 40GB IDE HDD
  - Dual Gigabit Ethernet (GbE) controllers
  - Double-wide blade (uses 2 BladeCenter slots)
  - Infiniband (IB) Option:
    - Qty 0-2 IB 4x Host Channel Adapters
- BC Chassis Configuration **(~3.5TFLOPS peak)**
  - Standard **IBM** BladeCenter One
  - Max. 7 Blades per chassis (QS20 - 2 slots each)
  - 2 Gigabit Ethernet switches
  - External IB switches required for IB option



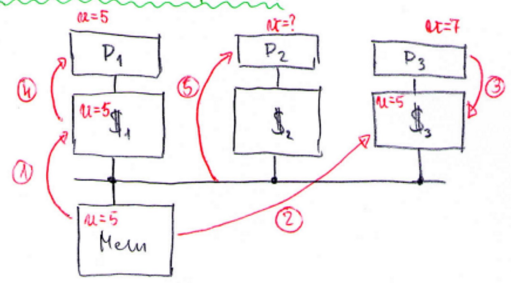
Cell Blade



Cell BE dual processor blade:
- → dual Cell BE processors
- → 2 × 512 MB DRAM
- → 40 GB HDD
- → dual gigabit ethernet controllers
- → PCI express connection interface

6. Show the cache coherence problem in a bus based multiprocessor! (3p)

Cache coherence problem



① P_1 reads u, $u=5$ from memory
② P_3 reads u, $u=5$ from memory
③ P_3 writes u, $u=7$ put into $\$_3$
④ P_1 reads u, since it is in cache (see ①) its value is from $\$_1$, which stores $u=5$ ↯
⑤ P_2 reads u, but what is the value of it?

7. Specify the main feattures of the GPU architecture (block diagram , programming) (3p)

GPU architecture - graphics pipeline

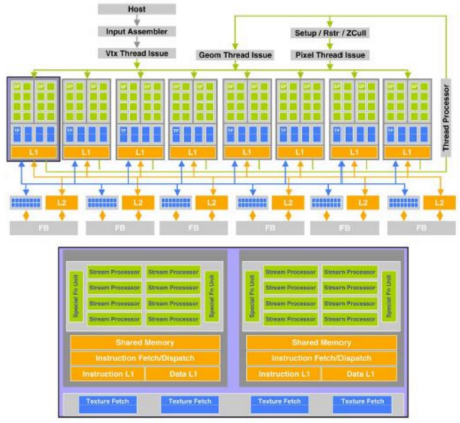
- Input is list of geometric primitives, typically triangles. Specific steps:
- Vertex operations – transformed into screen space and shaded, lights, parallel
- Primitive assembly – vertices are assembled into triangles
- Rasterization – determining which screen-space pixel locations are covered by each triangle, each triangle generates a fragment
- Fragment operations – color, textures, parallel
- Composition – assembling the final image, closest fragment to the camera

Evolution of GPU architecture

- The key step was replacing the fixed-function per-vertex and per-fragment operations with user-specified programs run on each vertex and fragment.
- Current GPUs support the unified Shader Model 4.0: supports at least 65k static instructions and unlimited dynamic instructions/ 32-bit integers and floating-point numbers/ arbitrary number of direct and indirect reads from global memory/ loops and branches

Architecture of a modern GPU

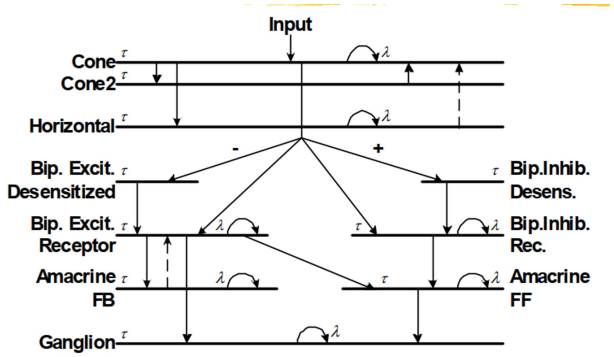
- CPU divides the pipeline in time – ~20 cycles/ operation
- GPU divides the pipeline in space, the part of the processor working on one stage feeds its output into a different part that works on the next stage – thousands cycles/operation
- Successful because:
  - In any stage could exploit data parallelism
  - Each stage's hardware could be customized with special-purpose hardware for its given task



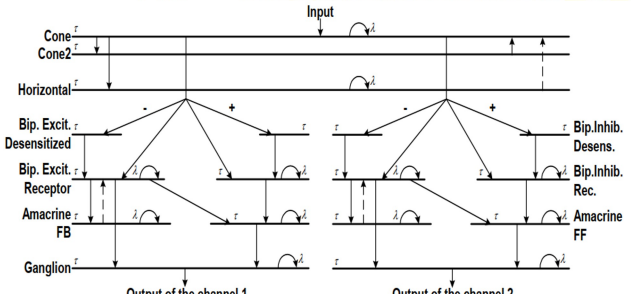
Architecture of a modern GPU

- Disadvantage: GPU pipeline is dependent on its slowest stage
- Strict pipelined task-parallel architecture -> increasingly built around a single unified data-parallel programmable unit
- Benefit: with all programmable power in a single hardware unit

8. Specify the architecture, the main parameters of a retina model (3p)



Simlified structure of the two channel retina model



Real-time emulation

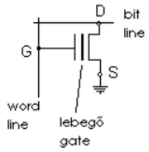
- Real time emulation of two retina channel (25fps, 30bit accuracy)
  - Timestep 2^{-7}ms
  - 128 000 iteration/s
  - 64x64 sized image
  - ~0.5 billion cell iteration/s
- XUPV2P30
  - 1 processor core (2 channel)
  - 190 MHz clock frequency
  - 190 million cell iteration/s
  - 32x32 sized image
- Virtex-II 3000
  - 1 processor core (2 channel)
  - 133 MHz clock frequency
  - 133 million cell iteration/s
  - 32x32 sized image

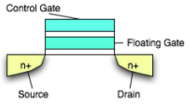
1. Please describe the floating gate FPGA programming method (advantages/disadvantages). (2p)

d.) Floating gate

Double gate technology is used, one of the gates are floating gates.The other gate is fixed gate. INTEL is using this method.

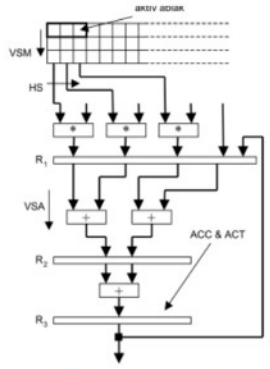
- Properties:
  - Programing : charge will be delivered to floating gate, the transistor will be open,
  - writing: charge will be deliverd to the floating gate by using larger potential.
  - You can delete the progarams as well by using UV light
  - The program is saved even if the power off.
  - large 2-4 kΩ resistance
  - large 10-20 femtoF parazitic capacitance



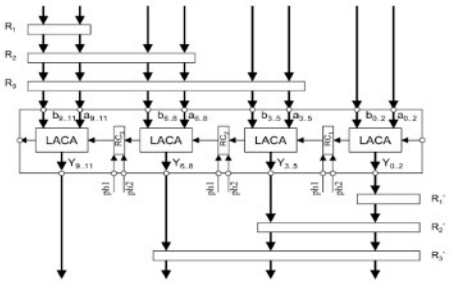


2. Specify the types of wires in an FPGA, how can be minimized the wire delay on an FPGA? .(3p)

3. Specify the pipeline operation of the CASTLE CNN (block diagram) (3p)



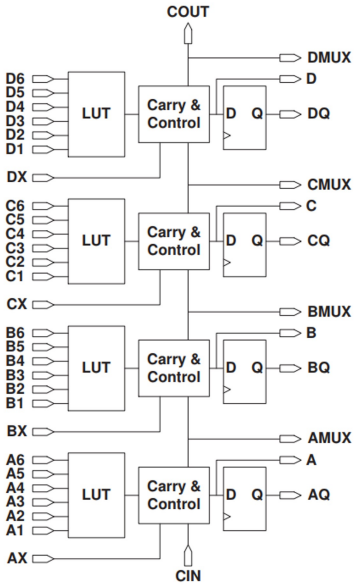
CASTLE architecture with pipeline between the multipliers and adders.



Pipeline inside the multipliers and adders

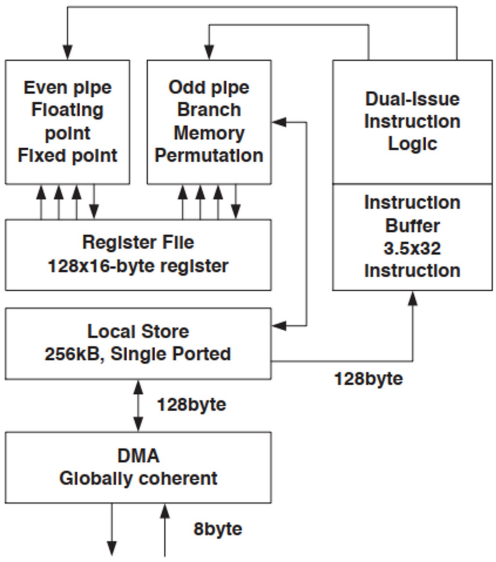
5. Milyen megfontolásokkal csökkenthető a kivezetés-szám igény az emulált digitális CASTLE architektúrában? (3p)

4. Draw a Xilinx Virtex 5 FPGA Slice (3p)



5.13. ábra. A Virtex-5 slice szerkezete.

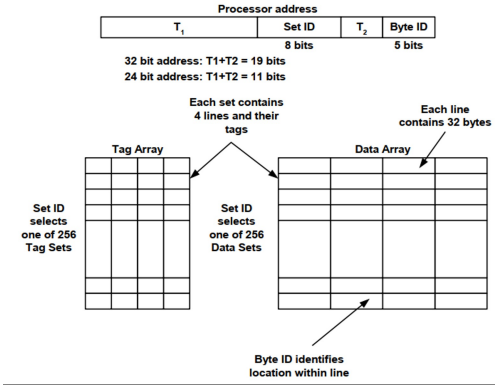
5. Draw the block diagram of a Synergic Processing element of an IBM Cell processor! (3p)



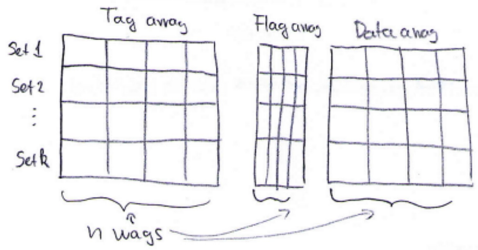
4.3. ábra. Block diagram of the Synergistic Processor Element

6. Draw the block diagram of a 4-way set associative cache memory with the following parameters: cache size: 64kbyte, cache line size: 32byte, address bus width: 32bit! (3p)

Organization of a 4-way set associative cache



Set associative cache



A memory address is split into

Tag ID	Set ID	Byte ID

Byte ID: position of the byte in one element of the data array.

Eg. if Data array = set of 32 B words, then byte ID: 0-31 ⇒ 2^5 bit describes it.

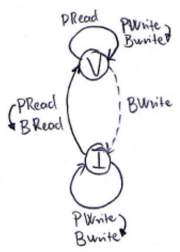
Set ID: identifies the set in which the Tag ID is possibly stored. length: $\log_2(k)$, so, Eg. for 256 sets ⇒ 8.

Tag ID: tag for the data. Length is the remaining. Eg. for 40 bit architecture the previous examples used 5 bits for Byte ID, 8 bits for Set ID ⇒ 40-8-5 = 27 bit Tag ID.

7. How does the MSI cache coherence protocol work? Show it through an example! (3p)

7. Mutassa be egy példán keresztül a cache koherencia problémát! Mutasson rá egy egyszerű megoldást (snooping bus) (3p)

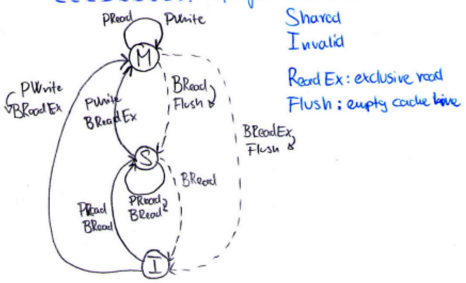
Snoop Bus — 2 flags: Valid, Invalid



P, B: processor, bus
→ processor initiated
--> bus-snooper initiated

0.) First, every line of the cache is Invalid.
1.) Data is read by the processor:
PRead ⇒ BRead and I⇒V
2.) Processor reads again
PRead ⇒ – and V⇒V
3.) Processor writes data
PWrite ⇒ BWrite V⇒V
and the write through changes the main memory
4.) Other processor writes data
BWrite snooped on bus, V⇒I

MSI protocol — 3 flags: Modified, Shared, Invalid



ReadEx: exclusive read
Flush: empty cache line

0.) V bit is Invalid
1.) Data is read by P if Invalid
PRead ⇒ BRead and I ⇒ S
2.) Data is written by P if Invalid OR Shared
PWrite ⇒ BReadEx and I/S ⇒ M
3.) If modified or shared data is read, the state won't change.
4.) Others initiate a Flush, if the data was is state M.

8. Give the explicit Euler form (advantages/disadvantages). (3p)

$\Delta t = h$ ekvidisztáns időlépés: $x(0), x(h), x(2h), \ldots$: h>0:

$x_k = x(t_k)$: $x_{k+1} = x(t_k + h)$:

$$\boxed{x_{k+1} = x_k + h\dot{x}_k = hf(x_k)}$$

explicit formula

Alapkérdés: Konvergens lesz ez az algoritmus? Sajnos nem mindig!

tehát nem lesz konvergens a közelített sorozat, sőt ***numerikusan divergens lesz egy stabil neurális áramkör digitális realizációja!*** De pontos!

9. Specify the architecture, the main parameters of a retina model. (3p)

Fenn

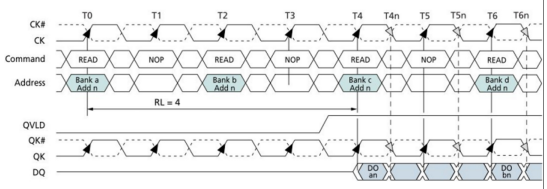
9. Milyen típusú párhuzamos architektúra a „space variant” CNN ? (2p)

4. Hogyan gyorsítható az emulált digitális retina modell szimulációja és milyen áron? (3p)

Kevesebb bit -> gyorsabb -> ára: kevesebb adatot tudsz mozgatni

8. Rajzolja fel egy RLDRAM memória írás/olvasási diagrammját! (3p)

RLDRAM 3 read timing



RLDRAM 3 write timing