Smart Instruction Codes For In-Memory Computing Architectures Compatible With Standard Sram Interfaces

Henri-Pierre Charles, Maha Kooli, Clément Touzet, Bastien Giraud, Jean-Philippe Noel

To cite this version:
Henri-Pierre Charles, Maha Kooli, Clément Touzet, Bastien Giraud, Jean-Philippe Noel. Smart Instruction Codes For In-Memory Computing Architectures Compatible With Standard Sram Interfaces. 2018. cea-01757665

HAL Id: cea-01757665
https://hal-cea.archives-ouvertes.fr/cea-01757665
Submitted on 3 Apr 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
SMART INSTRUCTION CODES FOR IN-MEMORY COMPUTING ARCHITECTURES COMPATIBLE WITH STANDARD SRAM INTERFACES

Maha Kooli, Henri-Pierre Charles, Clément Touzet, Bastien Giraud, Jean-Philippe Noel
Univ. Grenoble Alpes, CEA, LETI/LIST, FRANCE

DATE’18
Dresden, Germany, March 22nd, 2018
BREAK THE MEMORY WALL! BUT HOW...?

"memory wall" or "funnel effect"...

...is nowaday the main limitation for high performance computing
BOTTLENECK LIES IN THE MEMORY HIERARCHY

Memory access is **STILL** a bottleneck, even in GPUs…

Let’s do multi-core processors!

seems a good idea!

Source: Barcelona Supercomputing Center

Source: nVidia
GPU computing model (SPMD) need to:
- Copy/transfer data
- Group parallel instructions

IMPACT computing model:
- No copy / transfer
- Fine grain parallel scalar interleaving
BRING THE COMPUTATION INTO MEMORY

When data start to look like motorists in the *traffic jam* during the rush hour…

\[(\text{data}@\text{memory} \leftrightarrow \text{data}@\text{comp\_unit})\]

…it’s time to consider *teleworking*, in other word the *in-memory computing*
**Von Neumann Model:**
- Data & instruction in the same memory
- i.e. instructions are data
- SoC or PCB

**Memory INSN:** `ld r1 = @r2`
- 1 memory access (for the instruction)
- 1 instruction cycle (Decode + RF + memory access)

**Compute INSN:** `add r1 = r2 + r3`
- Compute instruction
- 1 memory access (for the instruction)
- 1 computation (Decode + RF + ALU)
The largest part of power consumption of logic and arithmetic operations is due to the memory access!

- The way to perform basic operations has to be restudied
- A lot of applications should be improved in performance
## DIFFERENT APPROACHES

<table>
<thead>
<tr>
<th>Process</th>
<th>Technique</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embedded Memories (CMOS Process)</td>
<td>In-Memory Computing</td>
<td>- Additional logic in memory</td>
</tr>
<tr>
<td></td>
<td>On-chip Memory</td>
<td>- Non-destructive computing</td>
</tr>
<tr>
<td></td>
<td>Additional Logic</td>
<td>- Non Volatile/Volatile Memories</td>
</tr>
<tr>
<td></td>
<td>SRAM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CPU</td>
<td>[Akyel’16] [Aga’17] [Kooli’17]</td>
</tr>
<tr>
<td>Stand-alone Memories (DRAM Process)</td>
<td>Logic-in-Memory</td>
<td>- Non-volatile Memories</td>
</tr>
<tr>
<td></td>
<td>On-chip Memory</td>
<td>(ReRAM, …)</td>
</tr>
<tr>
<td></td>
<td>Logic Operation</td>
<td>- Destructive computing</td>
</tr>
<tr>
<td></td>
<td>SRAM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CPU</td>
<td>[Matsunaga’09]</td>
</tr>
<tr>
<td></td>
<td>SRAM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CPU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Off-chip Memory</td>
<td>- Planar / 3D process</td>
</tr>
<tr>
<td></td>
<td>DRAM</td>
<td>- Non-destructive computing</td>
</tr>
<tr>
<td></td>
<td>CPU</td>
<td>[Gokhale’95] [Pugsley’14] UpMem</td>
</tr>
</tbody>
</table>
OUTLINE

• Introduction & Context

• In-Memory Power Aware CompuTing (IMPACT)

• IMPACT Memory Instruction Code

• IMPACT Communication Protocol

• Conclusion & Perspectives
In-Memory Power Aware CompuTing (IMPACT)
Computing in dedicated units:
- High data transfer between the ALU & the memory
  - Power hungry
  - Interconnect & memory security issues

In-memory computing:
- Reduced data transfer
  - Energy-efficient
  - Execution time acceleration
  - Security reinforcements (limitation of the side channel attacks (on buses))
IMPACT MEMORY

### SRAM architecture

<table>
<thead>
<tr>
<th>Row decoder</th>
<th>SRAM bitcell array</th>
<th>CTRL</th>
<th>IO</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN/OUT DFF/LATCH</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### IMPACT Memory

- **Multi-row selection**
- **IO to ALU-like**

<table>
<thead>
<tr>
<th>multi-row selector (&gt;2)</th>
<th>SRAM bitcell array</th>
<th>CTRL</th>
<th>ALU-like</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN/OUT DFF/LATCH</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Enable in-memory operations
  - Reduce latency & energy consumption due to data transfer
- SRAM bit cell array allows:
  - Long word arithmetic/logic operations not limited by register size, but with memory line size
  - Multi-row selection for some logic operations
  - Simultaneous storing in different addresses

---


---

**DATE’18 | Henri-Pierre Charles | 22/03/2018 | 12**
Emulate the IMPACT system features
- Long word operations
- Multi-operand operations

LLVM
- Early design stage of the system: Not defined ISA
- Manipulate arithmetic/logic operations on large vectors

Target Applications
- Image Processing (*Motion Detection*)
- Cryptography (*One Time Pad*)

Experimental Gains
- Execution time: up to 6145x
- Energy: up to 12,9x
IMPACT Memory Instruction Code
• Initial idea : put logic operation in bitcells - done
• Added idea : add parallel arithmetic in I/O - done
• Create an high level emulation platform (based on LLVM) - done
• New idea : create an « inverted Von Neuman » protocol (aka ISA) - (this presentation) done
• Tape out april 2018 - on going
• Create a more accurate emulation platform - on going
• Create compilation toolbox - on going
• Evaluate high level benchmarks - on going
• ../..
### Logic & Memory Operation

<table>
<thead>
<tr>
<th>Logic &amp; Memory Operation</th>
<th>Memory</th>
<th>Shift</th>
<th>Logic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not</td>
<td>Set</td>
<td>Xor</td>
</tr>
<tr>
<td></td>
<td>Reset</td>
<td>Shift Left</td>
<td>Nxor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shift Right</td>
<td>And</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Or</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Nor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Nand</td>
</tr>
</tbody>
</table>

### Arithmetic Operation

<table>
<thead>
<tr>
<th>Arithmetic Operation</th>
<th>Memory line size word</th>
<th>8 bits words</th>
<th>16 bits words</th>
<th>32 bits words</th>
<th>64 bits words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory line size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>word</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtraction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decrement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Bit Positions**

- **MSB (Most Significant Bit)**
- **LSB (Least Significant Bit)**

- **More than two input operands**
- **Maximum two input operands**

**IMPACT OPERATIONS AKA OPCODES**
• Multi-operand operations (logic/memory operations)
• Problematic: encoding all the operand addresses in the instruction requires large bus size
  ➢ Propose a novel concept based on pattern construction

**IMPACT Instruction**

<table>
<thead>
<tr>
<th>Opcode</th>
<th>Pattern Code</th>
<th>Address</th>
<th>Mask</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR</td>
<td>0110</td>
<td>1100</td>
<td>1</td>
</tr>
</tbody>
</table>

**IMPACT Memory**

Row Selector | Pattern Register | SRAM Array (N-columns, 16-rows)
---|---|---
0 | 0 0 1 1 0 1 0 0 0 0 1 1 0 0 0 0 | 
1 | 1 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 | 
1 | 1 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 | 
1 | 1 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 | 
1 | 1 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 | 
1 | 1 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 | 

@: Mask: 0 1 1 0
The proposed method allows:

- Building regular patterns
- Patterns can be refined by adding/deleting a specific line
- Patterns can be stored in the pattern register for future use
- Selecting multiple lines in the SRAM array to perform the multi-operand operation
The conventional format of instruction with maximum two source addresses
Long-word operations (logic/arithmetic operations)

**TOW-OPERAND INSTRUCTION FORMAT**

<table>
<thead>
<tr>
<th>Opcode</th>
<th>Address 1</th>
<th>Address 2</th>
<th>SP</th>
<th>Output</th>
<th>SI</th>
</tr>
</thead>
</table>

- **add, sub, …**
- @ of 1\(^{st}\) & 2\(^{nd}\) operand
- @ where the result is stored

**A select pattern** bit to enable/disable the pattern construction using the row selector

**A smart instruction** bit:
- 0: if conventional instr.
- 1: if IMPACT instr.
IMPACT Communication Protocol
- Communicate in-memory instructions via **data & address** busses of a conventional system
  - Compatible with **existing** system **architecture** (*conventional system bus*)
  - Enable **interleaving** the **CPU** & the **in-memory** instruction execution

**DATE’18 | Henri-Pierre Charles | 22/03/2018 | 21**
1. Address the SRAM in conventional mode
2. Address the IMPACT memory for read/write
3. Address the IMPACT memory for computation

In-Memory Computing System

Data Bus:
- Opcode: 7-bits
- @1: 12-bits
- @2: 12-bits
- SP: 1-bit
- Total: 32-bits

Address Bus:
- SI=1
- @IMPACT: 1-bit
- @Output: M-bits
- MSB: [16-bits]
- Total: 32-bits

Instruction/Data Memories:
- SRAM
- Total: 4k words x 2^M*32-bits

In-Memory Computing System

Patent filed in December 2017
1. **Interleave CPU & IMC instruction execution**
   - Perform massive data computation inside IMC, and not optimized computation in CPU
     - For image **qqVGA 160x120 (not pipelined):**
       - Execution Time Speed-Up: **1376x**
       - Energy Reduction Factor: **29x**

2. **Perform all the computation inside IMC**
   - For image **qqVGA 160x120 (not pipelined):**
     - Execution Time Speed-Up: **765x**
     - Energy Reduction Factor: **29x**

### Imc Application

```c
typedef unsigned char ImgLine_attribute_(ext_vector_type(N));
typedef ImgLine Image[M];

for (line = 0; line < M; ++line){
    Img[line] = imgA[line]-imgB[line];
}
```

### DATA INTERLEAVING: A TEST CASE

**Execution Trace**

<table>
<thead>
<tr>
<th>Cycle</th>
<th>Instruction</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>sub 512</td>
<td>0x7d40</td>
</tr>
<tr>
<td>1</td>
<td>add 32</td>
<td>0x0001</td>
</tr>
<tr>
<td>2</td>
<td>sub 512</td>
<td>0x7e40</td>
</tr>
<tr>
<td>3</td>
<td>add 32</td>
<td>0x0001</td>
</tr>
<tr>
<td>4</td>
<td>sub 512</td>
<td>0x7f40</td>
</tr>
<tr>
<td>5</td>
<td>add 32</td>
<td>0x0001</td>
</tr>
<tr>
<td>6</td>
<td>sub 512</td>
<td>0x8040</td>
</tr>
<tr>
<td>7</td>
<td>add 32</td>
<td>0x0001</td>
</tr>
<tr>
<td>8</td>
<td>sub 512</td>
<td>0x8140</td>
</tr>
<tr>
<td>9</td>
<td>add 32</td>
<td>0x0001</td>
</tr>
<tr>
<td>10</td>
<td>sub 512</td>
<td>0x8240</td>
</tr>
<tr>
<td>11</td>
<td>add 32</td>
<td>0x0001</td>
</tr>
<tr>
<td>12</td>
<td>sub 512</td>
<td>0x8340</td>
</tr>
</tbody>
</table>

**CPU (32bit)**

- load i
- load M
- sub
- add
- sub
- add
- sub

**IMC (512bit)**

- sub
- add
- sub
- add
- sub
- add
- sub
Source Code

... 
R = s1 + s2 
...

Assembly Code

... 
mv r2, @R 
mv r1, #add 
shl r1, #6 
xor r1, #@s1 
shl r1, #6 
xor r1, #@s2 
store r1, r2 
...

Compilation

IMC Instruction:

add @s1 @s2 @R

Do not change the actual ISA

Overhead preparation

Architecture scenario w/o ISA modification
Conclusion & Perspectives
CONCLUSION & PERSPECTIVES

• Propose a new communication protocol between the CPU and the IMPACT memory
  ➢ Compatible with existing system architecture (conventional system bus)
  ➢ Enable interleaving the CPU & the in-memory instruction execution

• Work on the compiler:
  ➢ Generate the assembly code respecting the communication protocol
  ➢ Interleave the IMC & CPU instruction execution
    • Based on the performance evaluation
  ➢ Optimize the data set-up in the memory
    • Data alignment in the IMC
    • Data interleaving