How to cope with the Power Wall

Power consumption by internet: x30 till 2030 if trends continue

Power-Efficient Computing
- Power-efficient Microchip Design
- Power-efficient Computer Architectures
- Power-efficient Languages and Compilers
- Power-efficient Software Implementation
- Power-efficient Machine Paradigm

>> Outline <<
- The Power Wall
- "Dataflow" Computing
- Reconfigurable Computing
- Time to Space Mapping
- The Xputer Paradigm
- Conclusions

© 2015, reiner@hartenstein.de
http://xputer.de

Reiner Hartenstein (invited presentation): How to cope with the Power Wall; PATMOS 2015, the 25th International Workshop on Power and Timing Modeling, Optimization and Simulation; Salvador, Bahia, Brazil, September 1-5, 2015
The Power Wall

The Xputer Paradigm

Reconfigurable Computing

Time to Space Mapping

Crypto

“Dataflow” Computing

Optimization and

PATMOS 2015, the 25th International Workshop on Power and Timing Modeling, Optimization and Simulation; Salvador, Bahia, Brazil, September 1-5, 2015
What about Acceleration by Graphics Processors?

- Drastically smaller Speed-ups if at all
- Power saving mostly not documented
- R. Vaduc et al.: “... adding a GPU is equivalent to adding one more multicore CPU socket ...”

On the Limits of GPU Acceleration

Purpose

“The Design of a Special Purpose VLSI Chips...” M. J. Foster and H. T. Kung: The Design of Special-Purpose VLSI Chips — IEEE 7th ISCA, La Baule, France, May 6-8, 1980

1980 The Systolic Arrays (1)
no instruction streams needed

The super-systolic array: a generalization of the systolic array:
... now a general purpose methodology!

What Synthesis Method?

H.T.Kung: “of course algebraic!” (linear projection)

only linear pipes supports only very special applications with strictly regular data dependencies

The super-systolic array: a generalization of the systolic array: ... now a general purpose methodology!

My student Rainer Kress replaced it by simulated annealing; this supports also any irregular & wide shape pipe networks

What about Acceleration by Graphics Processors?

- Drastically smaller Speed-ups if at all
- Power saving mostly not documented
- R. Vaduc et al.: “... adding a GPU is equivalent to adding one more multicore CPU socket ...”

On the Limits of GPU Acceleration

Purpose

“The Design of a Special Purpose VLSI Chips...” M. J. Foster and H. T. Kung: The Design of Special-Purpose VLSI Chips — IEEE 7th ISCA, La Baule, France, May 6-8, 1980

1980 The Systolic Arrays (1)
no instruction streams needed

The super-systolic array: a generalization of the systolic array:
... now a general purpose methodology!

What Synthesis Method?

H.T.Kung: “of course algebraic!” (linear projection)

only linear pipes supports only very special applications with strictly regular data dependencies

The super-systolic array: a generalization of the systolic array: ... now a general purpose methodology!

My student Rainer Kress replaced it by simulated annealing; this supports also any irregular & wide shape pipe networks
Reiner Hartenstein, TU Kaiserslautern, Germany
http://hartenstein.de

Reiner Hartenstein (invited presentation): How to cope with the Power Wall; PATMOS 2015, the 25th International Workshop on Power and Timing Modeling, Optimization and Simulation; Salvador, Bahia, Brazil, September 1-5, 2015

H. T. Kung: “It’s not our job”
Who generates the datastreams?
without a sequencer: missed to invent a new machine paradigm
(The Xputer)

Duality of procedural Languages

Compilation: Software vs. Configware u. Flowware

Heterogeneous: Co-Compilation

© 2015, reiner@hartenstein.de
http://hartenstein.de

© 2015, reiner@hartenstein.de
http://xputer.de

TU Kaiserslautern

TU Kaiserslautern

TU Kaiserslautern

TU Kaiserslautern
Reiner Hartenstein (invited presentation): How to cope with the Power Wall; PATMOS 2015, the 25th International Workshop on Power and Timing Modeling, Optimization and Simulation; Salvador, Bahia, Brazil, September 1-5, 2015
Reiner Hartenstein, TU Kaiserslautern, Germany
http://hartenstein.de

Reiner Hartenstein (invited presentation): How to cope with the Power Wall; PATMOS 2015, the 25th International Workshop on Power and Timing Modeling, Optimization and Simulation; Salvador, Bahia, Brazil, September 1-5, 2015
The Reconfigurable Computing Paradox

although the effective integration density of FPGAs is by 4 orders of magnitude behind the Gordon Moore curve, because of:
more overhead
+ reconfigurability overhead
= etc.
Reinvent Computing
Enabling software developers to apply their skills over FPGAs has been a long and as of yet unreached research objective in reconfigurable computing.

Paradigm Shift Consequences

von Neumann: CPU
Program Counter (PC)
Configuration Xputer: Configure Engineering
Configuration Code (CC)
Data Counters (DCs) sequencing code (e.g. see MoPL language)
Xputer and vN: Heterogenous Engineering

Von Neumann Syndrome

Three conspiring tectonic shifts
- The energy constrained world
- From internet of people to internet of everything
- End of scaling as we know it

Shifting to the dominance of von-Neumann-only caused an cumulated damage of at least trillions of Dollars, if not quadrillions ....

Computing Paradigms

<table>
<thead>
<tr>
<th>Term</th>
<th>program counter</th>
<th>execution triggered by</th>
<th>paradigm</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>yes</td>
<td>instruction fetch</td>
<td>Instruction-stream-based (von Neumann)</td>
</tr>
<tr>
<td>DPU</td>
<td>no</td>
<td>data arrival**</td>
<td>Data-driven or data-stream-based Reconfigurable Computing</td>
</tr>
<tr>
<td>DFC</td>
<td>no</td>
<td>I-structure handling</td>
<td>Dataflow/Computer</td>
</tr>
</tbody>
</table>

*) based on tagged token "I-Structure"
**) "transport-triggered"
**MIT Tagged Token Dataflow Architecture**

I would call it

*Tagged Token Flow Architecture*

- no Program Counter
- no updateable global store

---

**I-Structures (I = incremental) - part 1**

-Jurij Silc: Dataflow Architectures

- **Problem:** Single-assignment role and complex data structures
  - each update of a data structure consumes the structure and the value producing a new data structure.
  - awkward or even impossible to implement.
- **Solution:** concept of I-structures
  - a data repository obeying the single-assignment rule
  - each element of the I-structure may be written only once but it may be read any number of times.
  - the status of each element of the I-structure can be:
    - present: the element can be read but not written,
    - absent: a read request has to be deferred but a write operation into this element is allowed,
    - waiting at least one read request of the element has been deferred.
  - The status of each element of the I-structure can be:
    - present: the element can be read but not written,
    - absent: a read request has to be deferred but a write operation into this element is allowed,
    - waiting at least one read request of the element has been deferred.
  - After data structures have become defined (a specific value assigned: can happen exactly once), all deferred reads, which are in the associated queue, become immediately satisfied.
- **I-structure makes it possible to use a data structure before it is fully defined.
- It allows defining complex data structures from existing though partially defined data structures.

---

**Power Efficiency of Programming Languages**

(an example)

---

**How big is big?**

- 10²⁴ Petabytes
- 10²³ Exabytes
- 10²² Zettabytes
- 10²¹ Yottabytes
- 10²⁰atto Bytes
- 10¹⁸ femto Bytes
- 10¹⁶ pico Bytes
- 10¹⁴ nano Bytes
- 10¹² micro Bytes
- 10¹⁰ nano Bytes
- 10⁸ byte

---

Reiner Hartenstein (invited presentation): How to cope with the Power Wall; PATMOS 2015, the 25th International Workshop on Power and Timing Modeling, Optimization and Simulation; Salvador, Bahia, Brazil, September 1-5, 2015