Computing for Data-Intensive Applications:

Beyond CMOS and Beyond Von-Neumann

Said Hamdioui

Computer Engineering
Delft University of Technology
The Netherlands

Workshop on Memristive systems for Space applications

European Space Agency, ESTEC
30 April 2015, Noordwijk, Netherlands

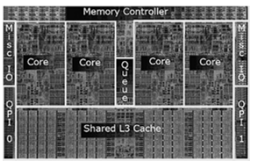
TU Delft
Delft University of Technology

The big picture

Applications

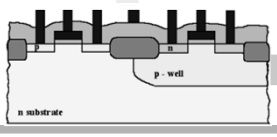
Storage
Computing efficiency

Computers



What is the solution?

Technology



- High cost
- Reduced reliability
- Saturated Clk
- Higher power

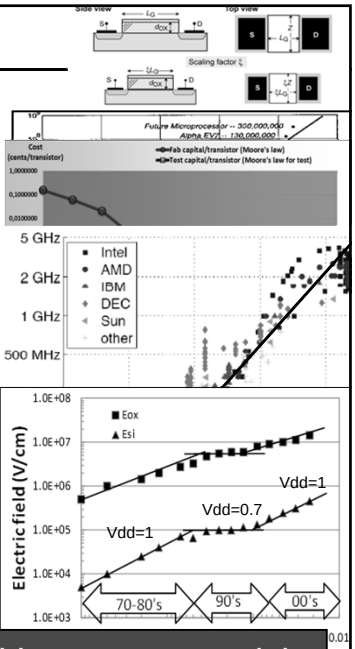
 2

Contents

- Technology
  - The good, the bad and the challenging
- Computing
  - The good, the bad and the challenging
- Future computers: possible scenarios
- Toward a new computing paradigm: **CIM Architecture**
  - Technology, potential & open questions
- Conclusion

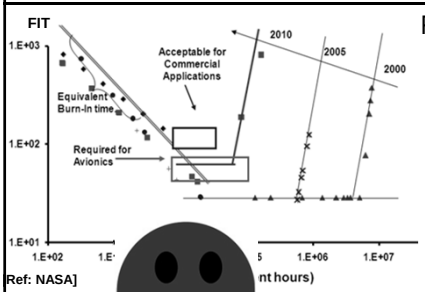
Technology…….The good

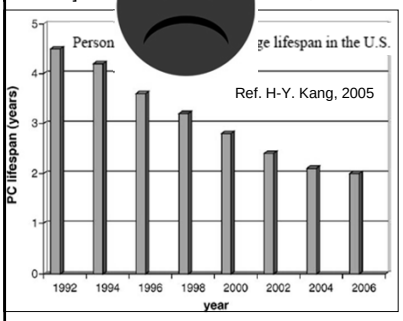
- Density: double
  - Dimensions reduce by 30%
    - WxL: 1x1 → 0.7x0.7 = **0.5**
  - Reduced IC cost
- Performance: 43% increase
  - Gate delay reduced by 30%
    - C=(0.7*0.7)/0.7= 0.7
    - freq =1/0.7 = **1.43**
  - uP freq ~ doubled every generation (till 2004)
- Power
  - Constant voltage scaling
    - Power= C x V^2 x f = **1**
  - Constant field scaling
    - Power= C x V^2 x f= **0.5**

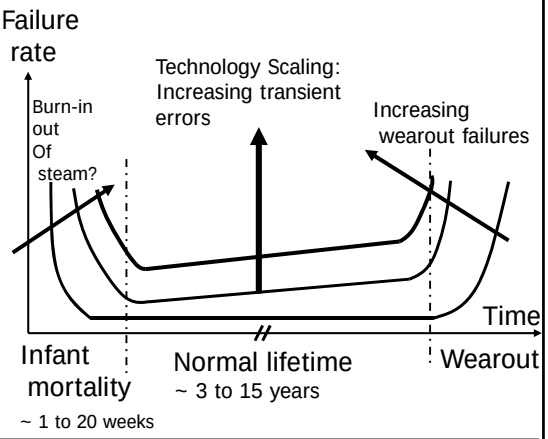


Scaling has been successful Financial investment model

Technology……. The bad





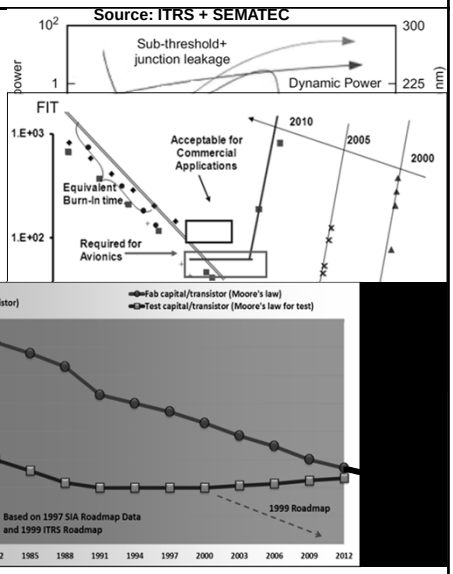


Higher failure rate & reduced life time

Components are becoming Unreliable

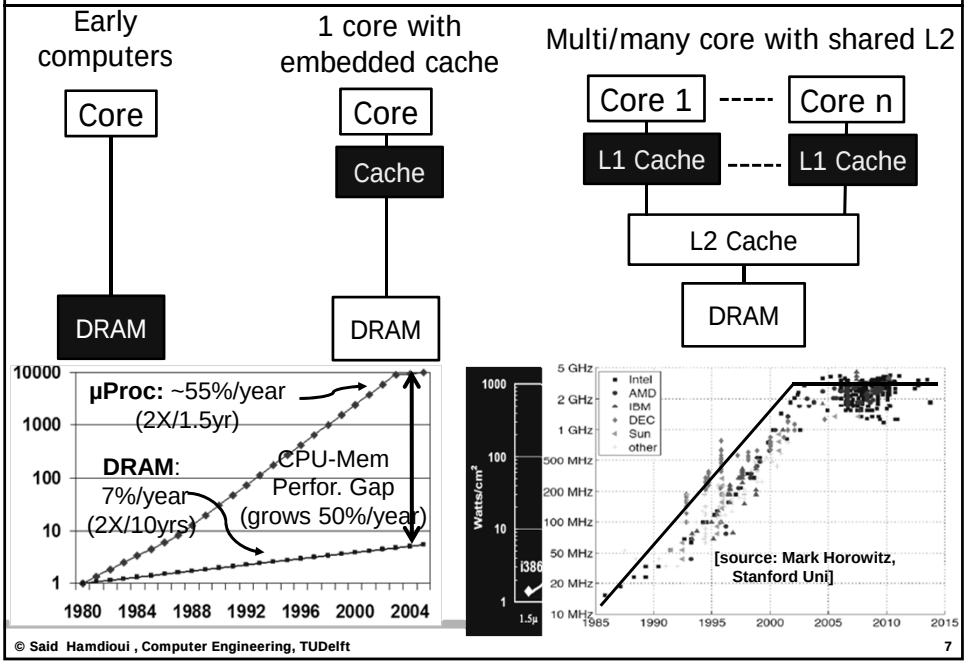
Technology……. The challenging (long term)

- High static leakage (volatile)
- Unreliable components
  - Expensive solutions
  - Economically not affordable
- Complex manufacturing
  - Low yield
  - High cost
  - Limited scalability
  - Comes at additional cost
- Less/no economical benefit



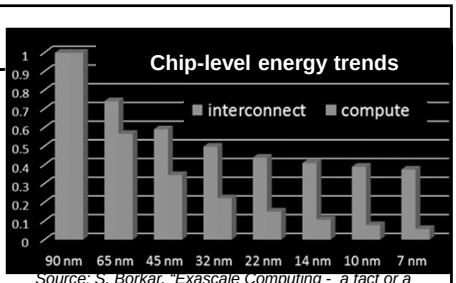
=> need of new device technologies

Computing…….The good



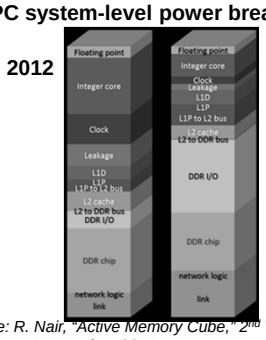
Computing…….The bad

- High energy consumption
  - Dominated by com & memory
  - 70 to 90% for data-ints. appl
- Limited data-bandwidth
  - Communication bottleneck
  - Stored program principle
- Complex programmability
  - Memory coherency
  - programmability overheard (scheduling policies, priorities, mapping, ..)
- Reduced / saturated performance
  - Enhancement based on expensive on chip memory (~70% of area)
  - Requires LD & ST: killers of overall perf



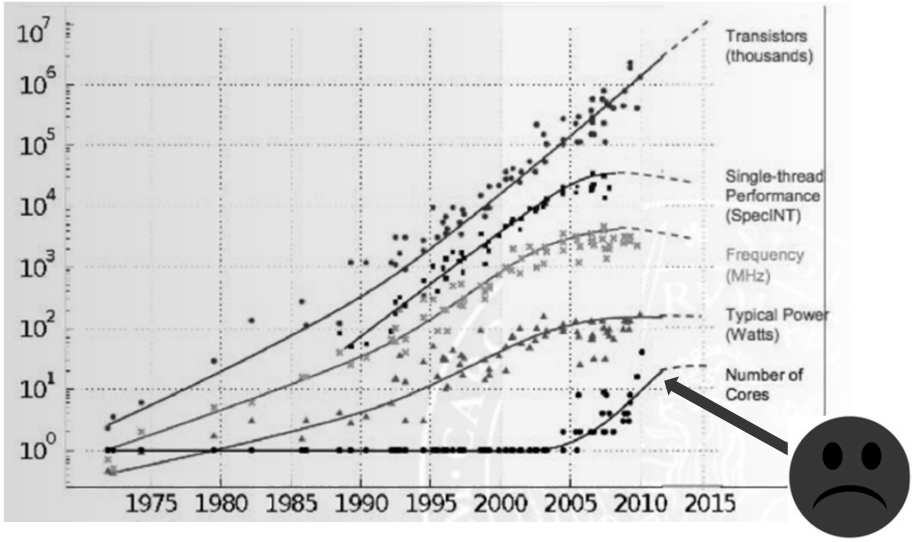
Source: S. Borkar, "Exascale Computing - a fact or a fiction?," IPDPS, 2013

HPC system-level power break-down



Source: R. Nair, "Active Memory Cube," 2nd Workshop on Near-Data Processing, 2014

Computing……. The bad

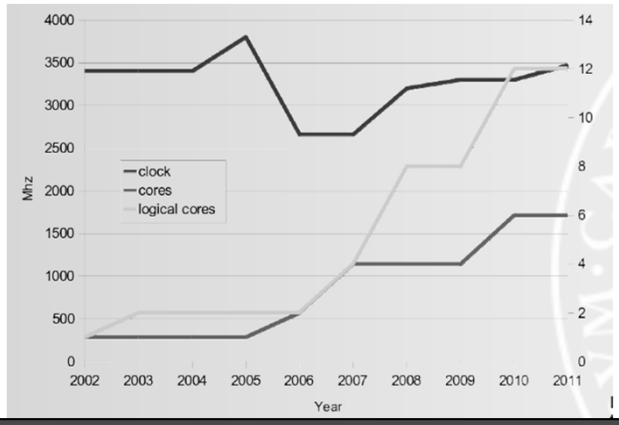


Ref: Jorn W. Janneck, Computing in the age of parallelism: Challenges and opportunities, Multicore Day, Sept 2013

TUDelft / Computer Engineering 9

Computing……. The bad

Intel desktop processor



Speed-up is no longer the result of a faster clock…
Instead, we get more **parallel devices**……
BUT: **hard to scale!!**

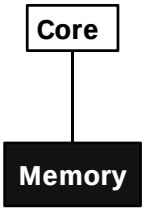
Ref: Jorn W. Janneck, Computing in the age of parallelism: Challenges and opportunities, Multicore Day, Sept 2013

TUDelft / Computer Engineering 10

Computing……. The challenging

- Extremely data intensive application (Big Data problems)
  - Economics, business, science, social media, healthcare, etc.
  - Data *storage* and *analysis*
  - E.g., 1PB@1000MB/sec = **12,5 days**!
    - Assume transfer rate of Front Side Bus at 1000MB/s
- Speed information increase exceed Moore's law
- Data size has already surpassed the capabilities of today's computation architectures
  - All suffer from communication and memory bottleneck

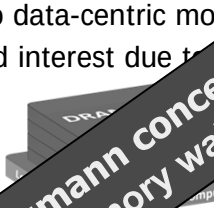
=> need of new architectures



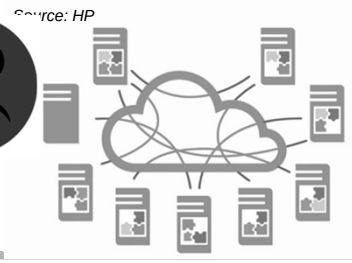
TUDelft / Computer Engineering 11

Future computers: possible scenarios

- Near memory computing
  - Bring computation closer to data
  - From computer-centric to data-centric model
  - Old concept, but renewed interest due t
    - Technology trends
    - Big data
    - New technologies (3D)
- HP: Machine
  - Discard con odel
  - Flatten ierarchie
  - Bring pro closer to the c
  - Electronics or computing, Photons for communications and ions for storage



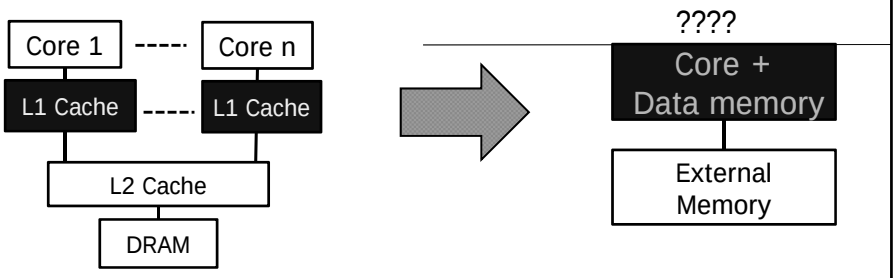
Still use Von Neumann concept
Communication/ memory wall



© Said Hamdioui , Computer Engineering, TUDelft 12

Toward a new computing paradigm

- How to address: memory wall/ communication and **big data**?

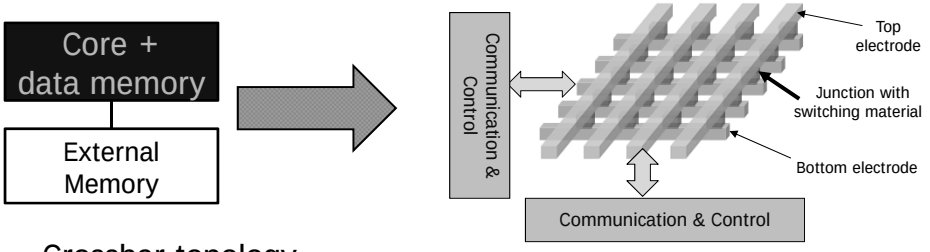


- Integrate *storage* and *computation* in the *same physical location*
  - Significantly reduces communication/ memory bottleneck
- Use non-volatile technology
  - Practically zero leakage
- Support massive parallelism
  - Crossbar architecture

[Source: S. Hamdioui, et.al, DATE 2015]

Toward a new computing paradigm

- **CIM architecture: Computing in Memory**

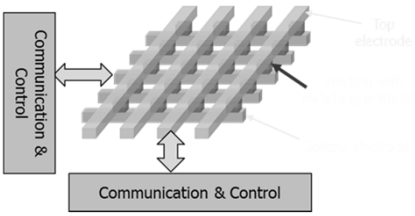


- Crossbar topology
  - Dense, non-volatile two terminal device at each junction
- No/ limited memory wall
  - Data will be loaded only at the first time
- Scalability at low cost
  - Significant reduction in area
  - Nanomtric device dimension (below 10 nm)

[Source: S. Hamdioui, et.al, DATE 2015]

Toward a new computing paradigm: CIM technology

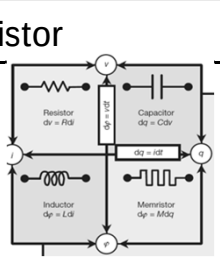
- Requirements for switching device
  - Dual functionality
    - Realize both memory and logic functions
  - Low energy consumption
    - Low/zero leakage: Non-volatility;
    - Reduce the overall power consumption
  - Scalability/ Nanometric dimensions
    - Realize extreme density at low price and reduce area
  - CMOS compatibility
    - Enhance manufacturing at low cost
  - Two terminal device with metal/insulator/metal structure
    - Realize the crossbar architecture
  - Good endurance & Good Reliability
- Solution
  - Emerging **resistive switching devices** = Memristors*

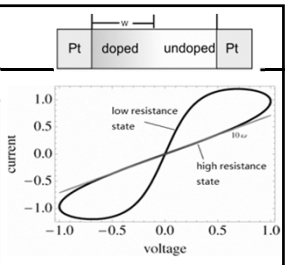


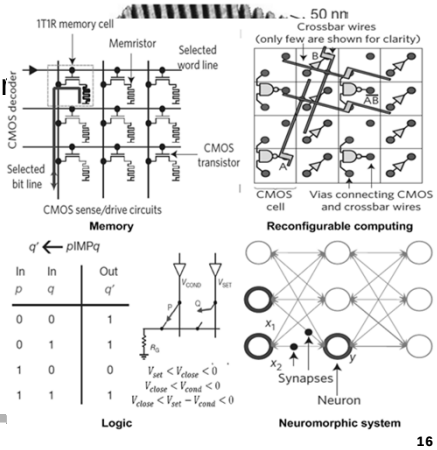
* L.Chua, "Resistance Switching Memories Are Memristors", 2014, Memristor Networks page 21-51, Springer

CIM technology: Memristor

- Invention
  - 1971: Leon Chua proposed
  - resistor with memory
  - "hysteretic" current-voltage characteristic
- Demonstration
  - 2008, HP manufactured memristor
  - Analysed different properties
    - integration density, leakage,..
- Potential applications
  - Memory
  - Logic
  - Etc







CIM potential

- Examples
  - Healthcare: DNA sequencing
    - we assume we have 200 GB of DNA data to be compared to
    - A healthy reference of 3GB for 50% coverage**

    *[**E. A. Worthey, Current Protocols in Human Genetics, 2001]*
  - Mathematic: 10^6 parallel additions
- Assumptions
  - Conventional architecture
    - FinFET 22nm multi-core implementation, with scalable number of clusters, each with 32 ALU (e.g comparator)
    - 64 clusters; each cluster share a 8KB L1 cache
  - CIM architecture
    - Memristor 5nm crossbar implementation
    - The crossbar size equals to total cache size of CMOS computer

[Source: S. Hamdioui, et.al, DATE 2015]

CIM potential

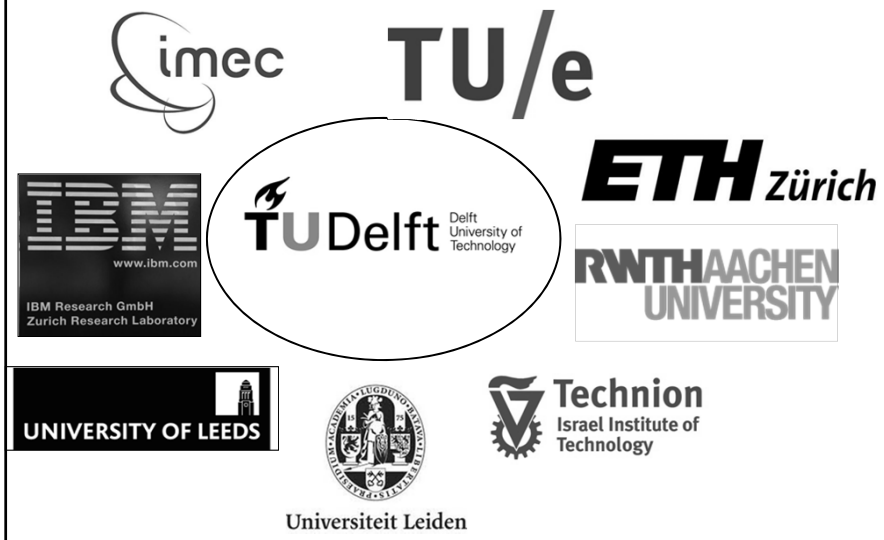
- Metrics
  - *Energy-delay/operation*
  - *Computing efficiency* : number of operations per required energy
  - *Performance area* : number of operations per required area
- Results

Metric	Archit.	DNA sequencing	10^6 additions	
Energy –Delay/ operations	Conv.	2.02e-03	1.5043e-18	> x100
	CIM	2.34e-06	9.25702-21	
Computing Efficiency	Conv.	4.11e01	6.5226e+9	> x100
	CIM	3.70e04	3.9063e+12	
Performance Area	Conv.	5.73e06	5.1118e+09	> x100
	CIM	8.28e09	4.9164e+12	

Key drives: Reduced memory bottleneck, non-volatile technology, massive parallelism

Hamdioui, et.al, DATE 2015]

CIM Partners



© Said Hamdioui , Computer Engineering, TUDelft 19

8. Conclusion

Do not only impact quality and reliability, but also YIELD!

- Constant voltage scaling (CMOS)
  - Complex faults & unreliable components
  - Higher leakage: volatile technology
  - Increasing cost: Lower yield, limited scalability, ...
  - Higher business pressure

=> **limits the applicability & benefitability**

Short term: Not only new tools/ IC flow required, but also innovations in DFX/BISX both for manufacturing and online/in field testing, monitoring, characterization, recovery, ...

Long term:
Alternative device technologies; E.g., Memristor

© Said Hamdioui , Computer Engineering, TUDelft 20

8.Conclusion

Do not only impact performance , but also energy!

- Von-Neumann based computers
  - Memory & communication bottleneck
  - Complex progammability of multi-cores
  - Higher power consumption
  - Big data & data-intensive applications

  => Unable to solve (today) and future big data problems

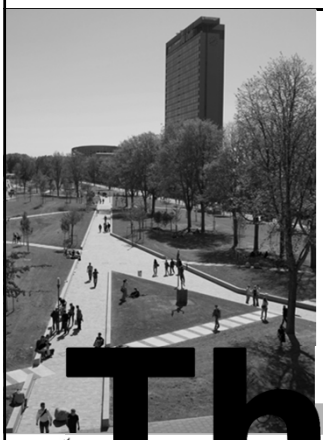
Short term:
- Specialization: application-specific accelerators (reduced prog)
- Near memory computing, accelerator around memories (data-centric model)

Long term:
Alternative architecture, beyond Von Neumann
E.g., CIM architecture

 21

Computing for Data-Intensive Applications:

Beyond CMOS and Beyond Von-Neumann



Said Hamdioui

Computer Engineering
Delft University of Technology
The Netherlands

Workshop on Memristive systems for Space applications

European Space Agency, ESTEC
30 April 2015, Noordwijk, Netherlands

Thanks

TU Delft
Delft University of Technology