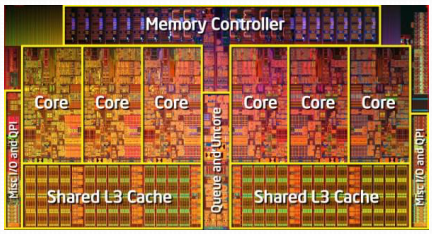
Energy Efficiency of Computing Architectures

Olivier Sentieys

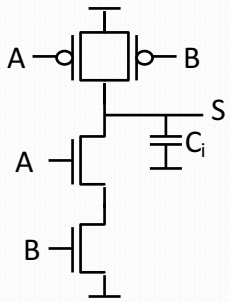
IRISA/INRIA – Cairn team
University of Rennes 1 olivier.sentieys@inria.fr

Power Consumption in Silicon Chips

- Chips, logic gates and transistors

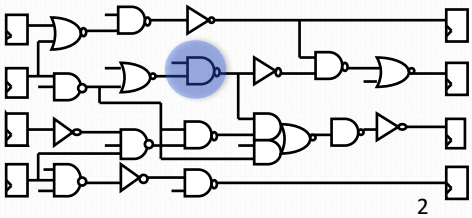


Intel's Xeon Chip



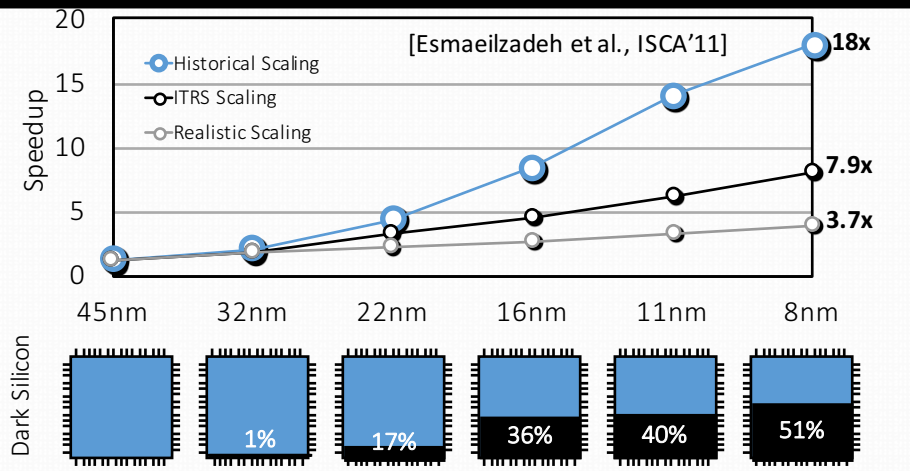
$$P_i = \alpha_i.f_i.C_i.Vdd^2 + I_{leak_i}.Vdd$$

$$P = \sum_i \left[\alpha_i.f_i.C_i.Vdd^2 + I_{leak_i}.Vdd\right]$$

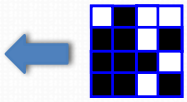


2

Multicore and Dark Silicon



- Replace dark cores with specialized cores (10-100x more energy efficient)



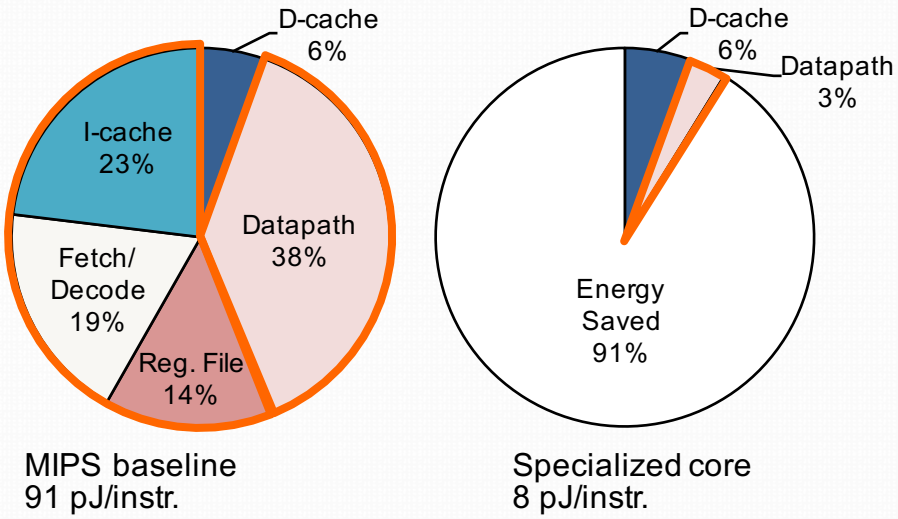
35

Outline

- Power and energy consumption basics
- Power consumption in processors
- Multicore: power and utilization walls
- Energy advantages of hardware accelerators
- Playing with accuracy for reducing energy
- Towards heterogeneous manycores
  - Can 3D stacking help?
  - Optical interconnect

36

Energy Savings in Specialized HW



MIPS baseline
91 pJ/instr.

Specialized core
8 pJ/instr.

[Goulding et al., Hot Chips'10]

37

An example: Bitcoin Mining

Type	Model	Mhash/s	Mhash/J	Power (W)
GPP	Intel Xeon X5355 (dual)	22.76	0.09	120
GPP	ARMCortex-A9	0.57	1.14	1.5
GPP	Intel Core i7 3930k	66.6	0.51	130
GPU	AMD 7970x3	2050	2.41	850
GPU	Nvidia GTX460	158	0.66	240
ASIC	AntMiner S1	180.000	**500**	360
ASIC	AntMiner S5	1.155.000	**1957**	590
FPGA	Bitcoin Dominator X5000	100	**14.7**	6.8
FPGA	Butterflylabs Mini Rig	25.200	**20.16**	1250

BITCOIN MINER

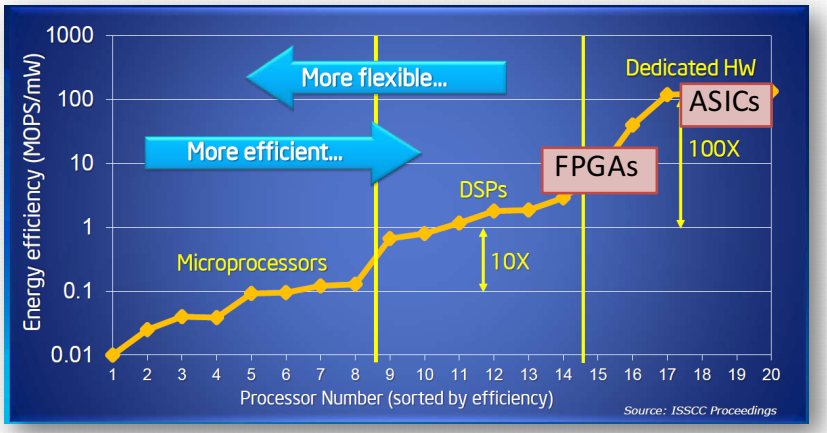
38

Time has Come for Specialization

- Microsoft Unveils Catapult to Accelerate Bing!
  - One FPGA per blade
  - 6 × 8 2-D torus topology
  - High-end Stratix V FPGAs
- Running Bing Kernels for feature extraction and machine learning
- Increase ranking throughput by 95% at comparable latency to software-only
- Increase power consumption by 10%
- Increase total cost of ownership by less than 30%

39

The Efficiency of Specialization



* Source: Ning Zhang and Bob Brodersen, ISSCC data

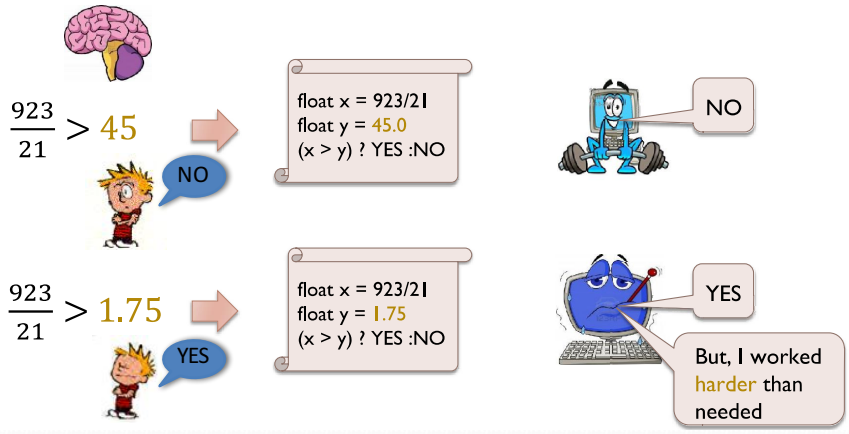
100-1000X Gap in Efficiency … but Specialization comes with Penalties in Programmability

Outline

- Power and energy consumption basics
- Power consumption in processors
- Multicore: power and utilization walls
- Energy advantages of hardware accelerators
- Playing with accuracy for reducing energy
- Towards heterogeneous manycores
  - Can 3D stacking help?
  - Optical interconnect

41

Humans Approximate.... But Computers Do Not!

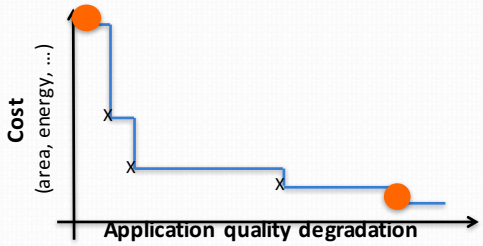


- Leads to inefficiency
- Overkill (for many applications)

42

Approximate Computing

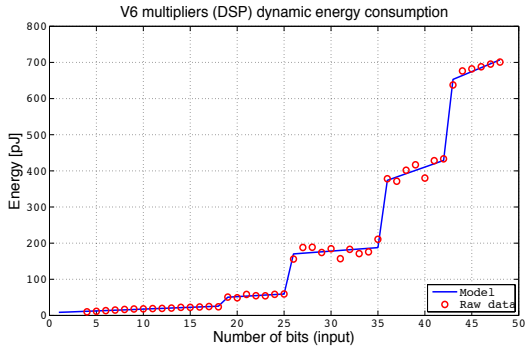
- Relies on the ability of many applications to tolerate some loss of quality
- Allows substantially improved energy efficiency by relaxing the need for fully precise operations
  - Trade quality against performance/energy



43

Energy vs. Data/Operator Size

- Arithmetic operators

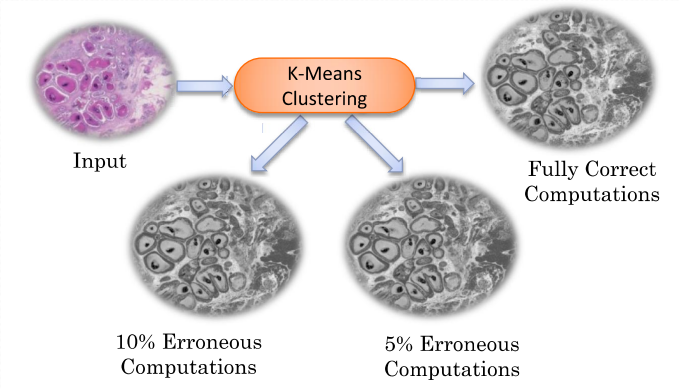


- Wire energy
  - 240fJ/bit/mm per transition
  - 32 bits, 10mm: 40pJ/word
  - 8 bits, 10mm: 10pJ/word
- Memory

44

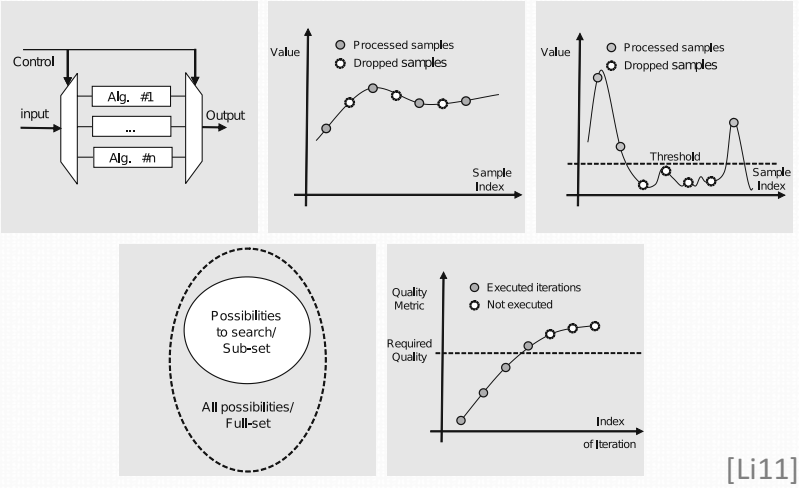
Intrinsic Application Resilience

- e.g. Image Segmentation



45

Algorithmic approximations

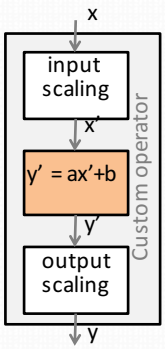


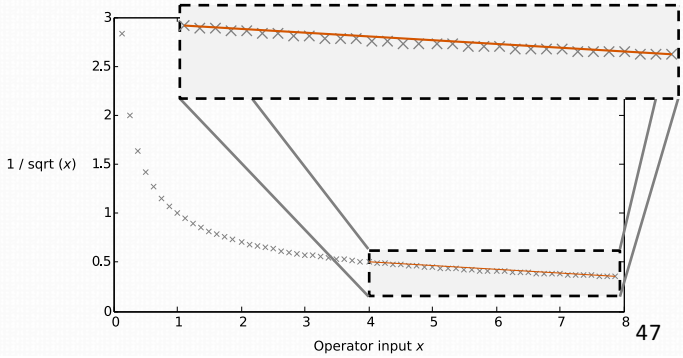
[Li11]

46

Algebraic approximations

- Real digital operators are not equal to mathematical operators
  - E.g. 1/sqrt(x) can be approximated by various techniques or polynomial orders
- Applications often accept even more aggressive approximations

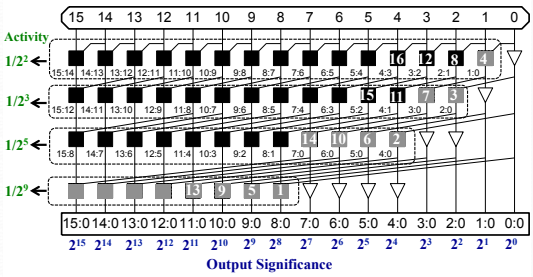




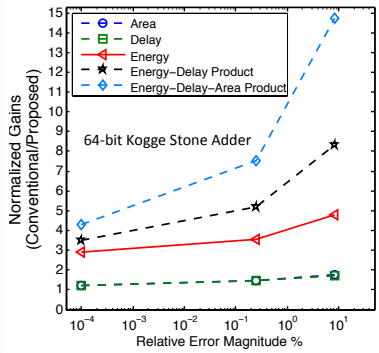
47

Approximate Operators

- Pruning logic cells
  - Han-Carlson and Kogge-Stone adders
  - Ranking metrics on logic gates



(a) Ranking Scheme for Minimizing Rel. Error and Average Error

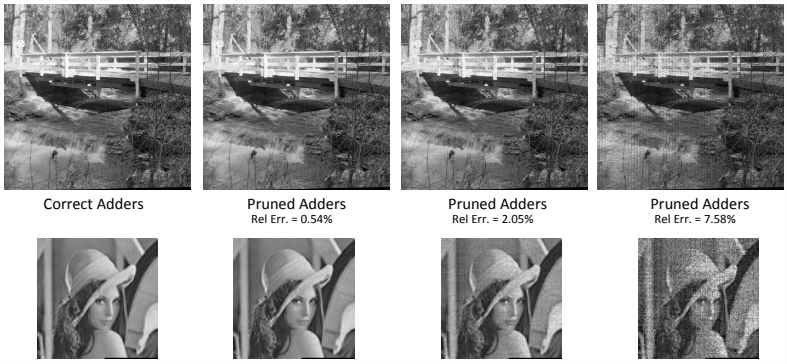


A. Lingamneni, C. Enz, J.-L. Nagel, K. Palem and C. Piguet, Energy Parsimonious Circuit Design through Probabilistic Pruning, DATE, 2011 (CF12, ACM TECS 2011)

48

Approximate Adders (CSEM/EPFL/RICE)

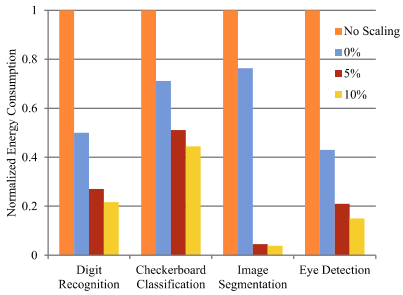
- Results after an FFT operation



49

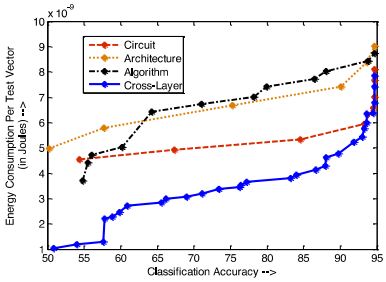
Approximate Recognition and Mining Processor

Energy Savings



- 1.3X-2X energy savings with no impact on output quality
- 2X-20X energy savings with 5% loss in quality

Benefits of cross-layer design (SVM Classification)



- Single level scaling does not exploit the inherent application resilience maximally
- Cross-layer optimization enables superior Energy-Quality tradeoff

V. K. Chippa, et. al, CICC 2013

50

Hard and Soft SIMD

Hard SIMD

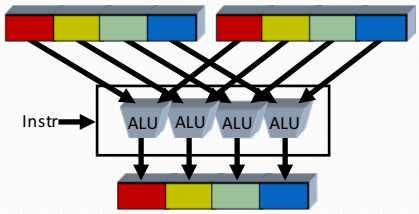
Needs HW modifications in the operators (e.g., manipulate the carry propagation of an adder)

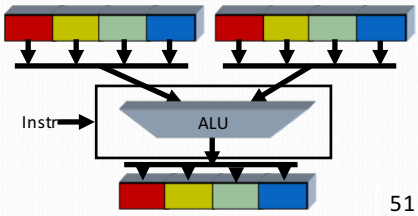
Very limited number of vector configurations (e.g., 1x64b, 2x32b, 4x16b, 8x8b)

Soft SIMD

The operator is unaware of the vector configuration

Large number of vector configurations (e.g., 5x12b, 6x10b, 1x32b+2x16b, etc.)





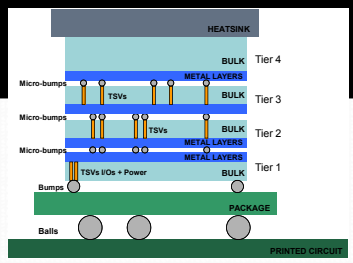
51

Outline

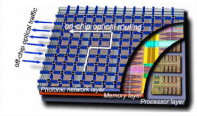
- Power and energy consumption basics
- Power consumption in processors
- Multicore: power and utilization walls
- Energy advantages of hardware accelerators
- Playing with accuracy for reducing energy
- Towards heterogeneous manycores
  - Can 3D stacking help?
  - Optical interconnect

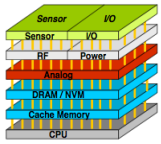
52

Chips go 3D!



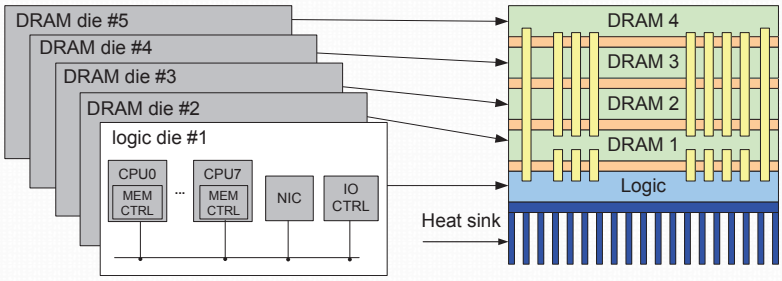
- 3D Integrated Circuits
  - Stack Multiple Dies
- Wire Length Reduction
  - Replace long, high capacitance wires by Through Silicon Vias (TSVs)
  - Low latency, low energy, high bandwidth
- Heterogeneous Integration
  - Image Sensors, Sensor Network Nodes
  - Processor + Memory





PicoServer

- Energy-efficient multicore architecture with 3D-stacked DRAM

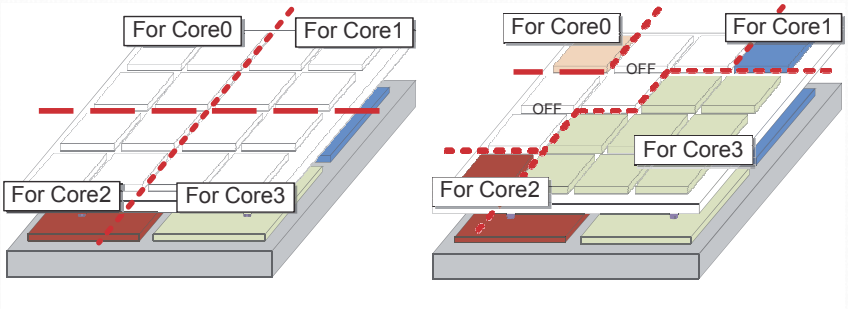


[Kgil et al., SIGOPS'08]

54

3D Memory Stacks

- Moving the compute closer to the data
- Non-Uniform Cache Architecture (NUCA)
  - Dynamic reconfiguration of cache structure

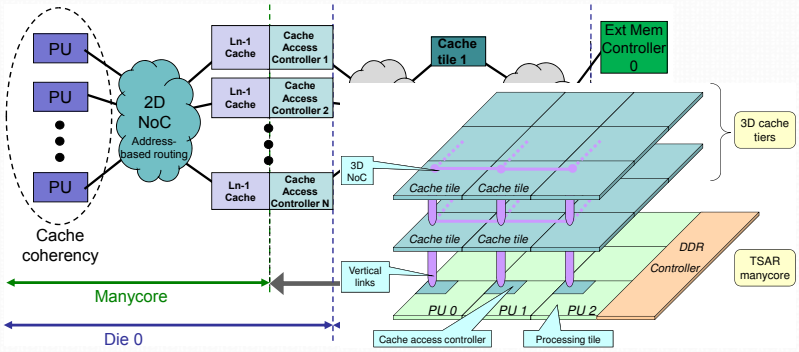


[Jung et al., GLSVLSI'11]

55

3D Memory Stacks

- Moving the compute closer to the data
- Non-Uniform Cache Architecture (NUCA)
  - Dynamic reconfiguration of cache structure



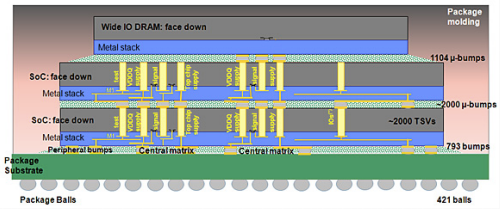
[Dutoit et al., DATE'13]

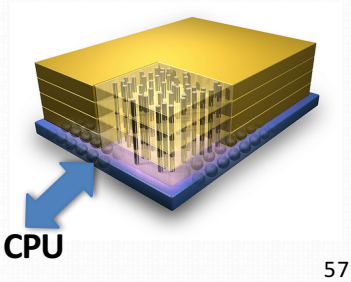
56

Wide I/O, Hybrid Memory Cube (HMC)

- Wide I/O memory interface
- 3D die stacks with TSVs
- 2.5D interposer
- 50GB/sec

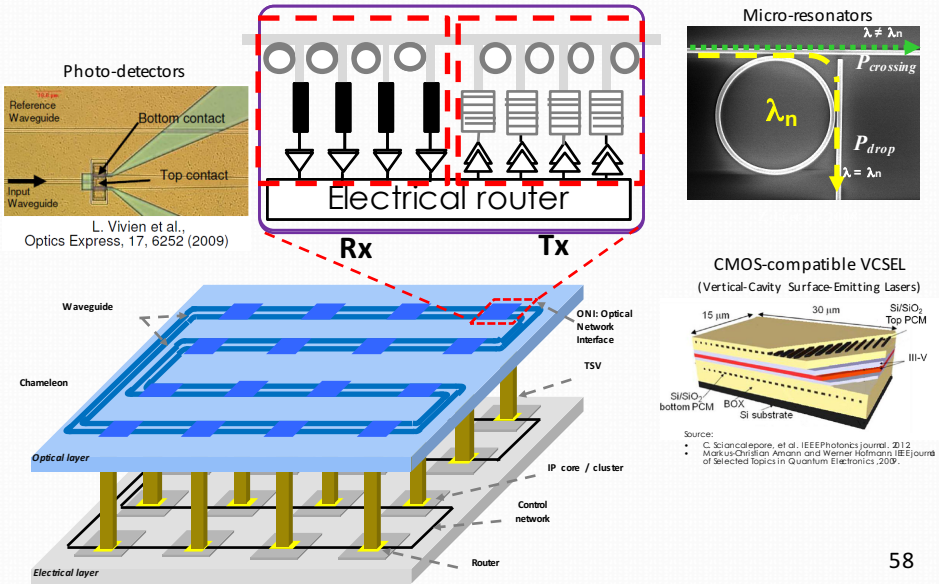
- Micron/Intel's HMC couples a logic layer with 3D-stacked DRAM
- 160GB/sec





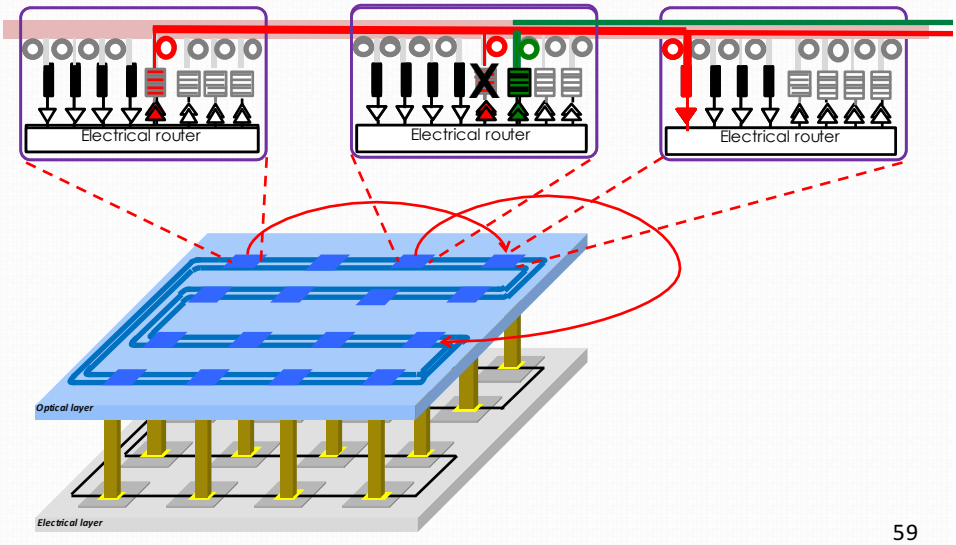
57

On-chip Optical Interconnects



58

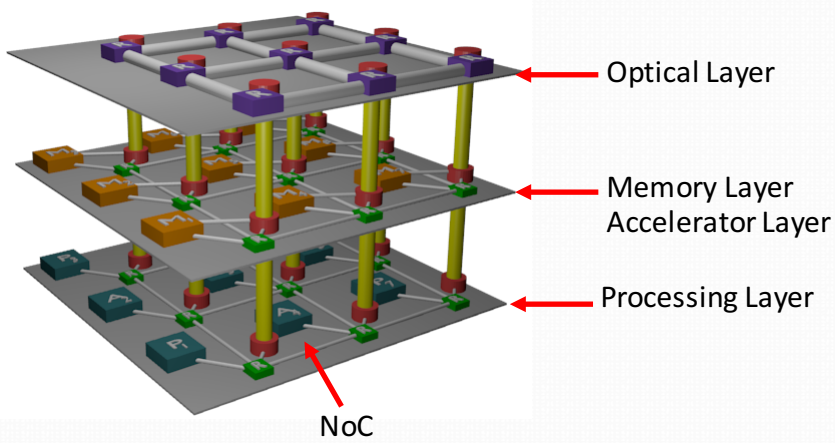
On-chip Optical Interconnects



59

3D Heterogeneous Multicores

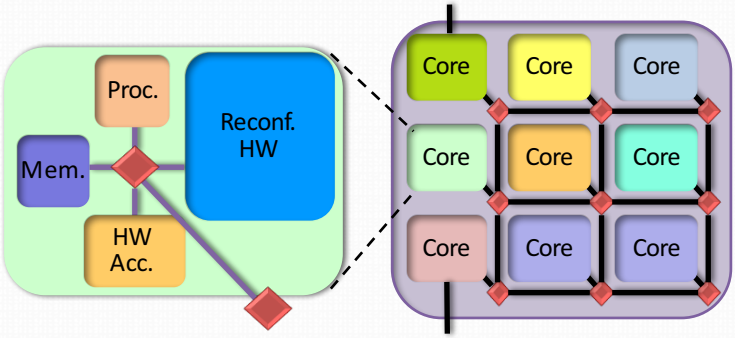
- 3D Optical Manycore Project



60

Towards Heterogeneous Multicores

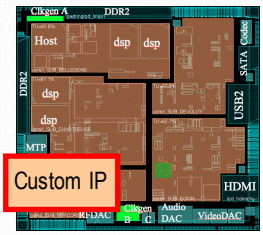
- Different cores on a single chip
  - GPPs, HW accelerators, memory, network-on-chip
- Self-adapting micro-architectures
  - Dynamically adapt the hardware to the application
  - Provide support for static and dynamic compilation



61

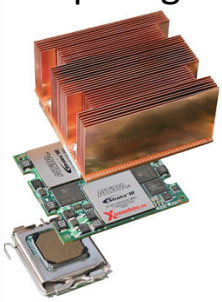
Towards Heterogeneous Multicores

- Embedded and High-Performance Computing



Embedded heterogeneous multicore

Heterogeneous platforms



FPGA accelerators for HPC

- C to hardware high-level synthesis boosts hardware designer productivity

62

Conclusions

- Energy consumption is a major issue
  - True in embedded systems since 20 years…
  - But also true now in HPC, mobile clouds, data centers, etc.
- What did we learn?
  - Hardware specialization
  - Inexact computation
  - Emerging technologies
- Dark Silicon is an opportunity
  - Heterogeneous manycore architectures
  - Bring a new demand for genuinely high level synthesis tools and (JIT) compilers that map programs to accelerators

63

References

1. Hadi Esmaeilzadeh, Emily Blem, Renee St. Amant, Karthikeyan Sankaralingam, and Doug Burger. 2011. Dark silicon and the end of multicore scaling. SIGARCH Comput. Archit. News 39, 3 (June 2011), 365-376.
2. Venkatesh, Ganesh, Sampson, Jack, Goulding, Nathan, Garcia, Saturnino, Bryksin, Vladyslav, Lugo-Martinez, Jose, Swanson, Steven, and Taylor, Michael Bedford, Conservation cores: reducing the energy of mature computations, Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2010.
3. William J. Dally, James Balfour, David Black-Shaffer, James Chen, R. Curtis Harting, Vishal Parikh, Jongsoo Park, David Sheffield "Efficient Embedded Computing" IEEE Computer, July 2008.
4. William J. Dally, keynote at IPDPS 2011.
5. Nathan Goulding, Jack Sampson, Ganesh Venkatesh, Saturnino Garcia, Joe Auricchio, Jonathan Babb, Michael Bedford Taylor, and Steven Swanson, GreenDroid: A Mobile Application Processor for a Future of Dark Silicon, Hot Chips 22, Stanford, CA, Aug. 2010.
6. Paul Chow, Why Put FPGAs in Your CPU Socket?, keynote at FPT 2013. http://www.fpt2013.org/Day3_keynote.pdf
7. Jongpil Jung, Kyungsu Kang and Chong-Min Kyung, Design and management of 3d-stacked NUCA cache for chip multiprocessors. *In Proc. of the 21st edition of the great lakes symposium on Great lakes symposium on VLSI*, GLSVLSI '11, pages 91–96, 2011. ISBN 978-1-4503-0667-6.
8. Denis Dutoit, Eric Guthmuller and Ivan Miro-Panades, 3d integration for power- efficient computing. *In Design, Automation Test in Europe Conference Exhibition (DATE), 2013*, mars 2013.
9. Taeho Kgil, Shaun D'Souza, Ali Saidi, Nathan Binkert, Ronald Dreslinski, Trevor Mudge, Steven Reinhardt and Krisztian Flautner, PicoServer: using 3D stacking technology to enable a compact energy efficient chip multiprocessor. *SIGOPS Oper. Syst. Rev.*, 40(5):117–128, octobre 2006. ISSN 0163-5980.

64

References

1. D.Menard, R.Rocher and O.Sentieys. Analytical Fixed-Point Accuracy Evaluation in Linear Time-Invariant Systems. IEEE Transactions on Circuits and Systems I: Regular Papers, 55(10):3197– 3208, November 2008.
2. K. Parashar, R. Rocher, D. Menard, O. Sentieys, D. Novo, and F. Catthoor. Fast performance evaluation of fixed-point systems with un-smooth operators. In Proc. of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD), pages 9–16, San Jose, CA, November 2010.
3. H.-N. Nguyen, D. Menard, and O. Sentieys. Dynamic precision scaling for low power wcdma receiver. In Proc. of the IEEE International Symposium on Circuits and Systems (ISCAS), pages 205–208, Taipei, Taiwan, May 2009.
4. http://blogs.msdn.com/b/satnam_singh/archive/2011/01/18/reconfigurable-data-processing-for-clouds.aspx
5. http://research.microsoft.com/en-us/projects/kiwi/default.aspx
6. David Greaves and Satnam Singh, Designing Application Specific Circuits with Concurrent C# Programs, in ACM/IEEE International Conference on Formal Methods and Models for Codesign, IEEE, 26 July 2010
7. http://researcher.watson.ibm.com/researcher/view_group.php?id=122
8. http://queue.acm.org/detail.cfm?id=2000516
9. Satnam Singh, Computing without Processors. Queue 9, 6, Pages 50 (June 2011), 14 pages. http://doi.acm.org/10.1145/1989748.2000516
10. Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors, Communications of the ACM, Vol. 54 No. 5, Pages 67-77, 10.1145/1941487.1941507