Domain-Specific Computing Platforms:
the Ultimate Energy-Efficiency of Hardware Accelerators

Olivier Sentieys

University of Rennes
IRISA/INRIA Rennes
sentieys@irisa.fr

CAIRN Project-Team
http://www.irisa.fr/cairn

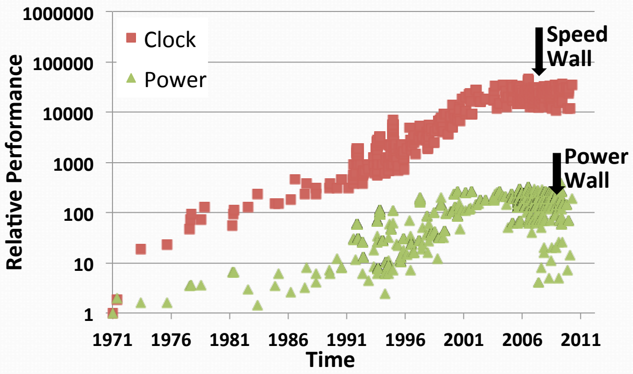
Motivations

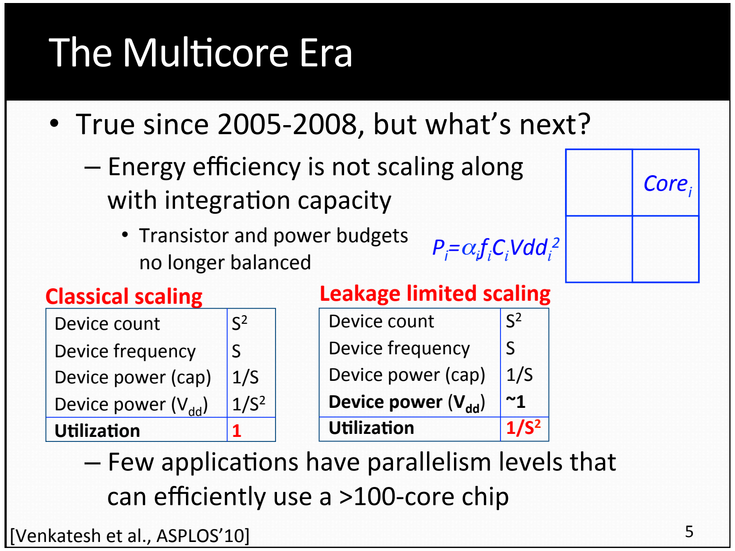
- A data center is not an embedded system!
  - But power is a major issue in ES since 20 years
- So what can we learn from embedded systems?
  - Hardware specialization
  - Adaptive hardware platforms
- Heterogeneous manycores
  - processors + accelerators + memory + network

Outline

- Multicore and the power wall
  - The Utilization Wall
  - Dark Silicon
- Energy advantages of hardware accelerators
- Reducing power on adaptive platforms
- Chips go 3D!
- Towards heterogeneous manycores

Limits Exist





The Multicore Era

- True since 2005-2008, but what's next?
  - – Energy efficiency is not scaling along with integration capacity
    - Transistor and power budgets no longer balanced

$P_i=\alpha_i f_i C_i Vdd_i^2$

Classical scaling

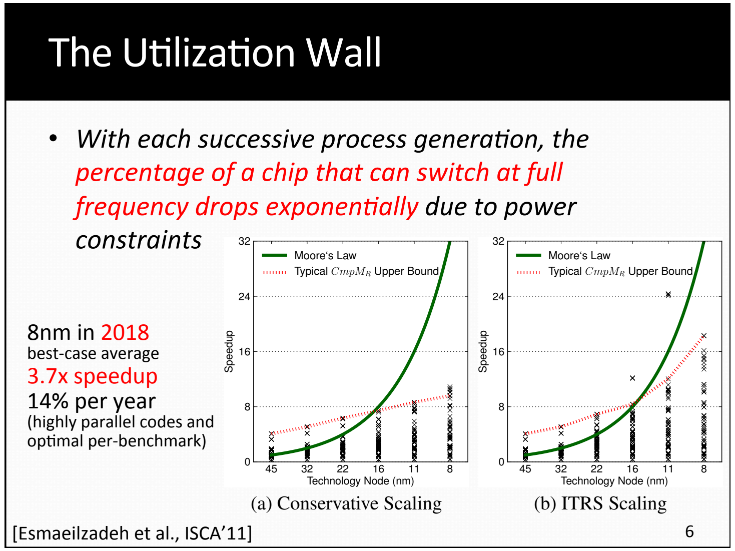
Device count	S^2
Device frequency	S
Device power (cap)	1/S
Device power (V_{dd})	$1/S^2$
Utilization	**1**

Leakage limited scaling

Device count	S^2
Device frequency	S
Device power (cap)	1/S
Device power (V_{dd})	**~1**
Utilization	**$1/S^2$**

- – Few applications have parallelism levels that can efficiently use a >100-core chip

[Venkatesh et al., ASPLOS'10]



The Utilization Wall

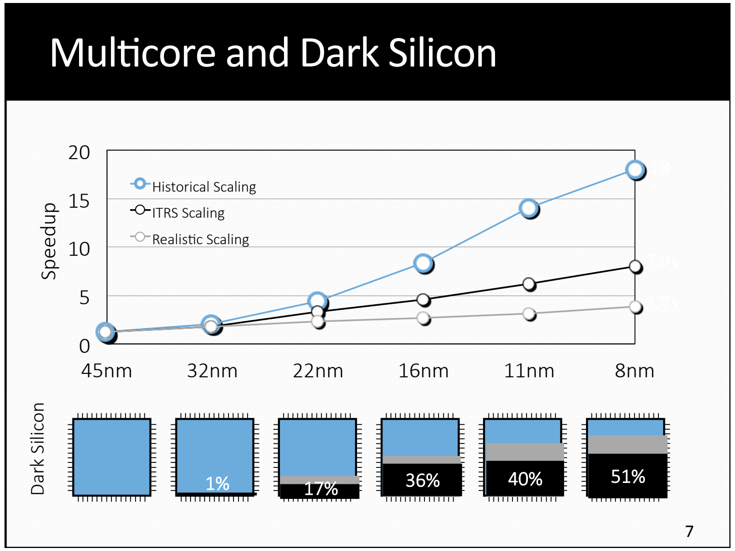
- *With each successive process generation, the percentage of a chip that can switch at full frequency drops exponentially due to power constraints*

8nm in 2018
best-case average
3.7x speedup
14% per year
(highly parallel codes and optimal per-benchmark)

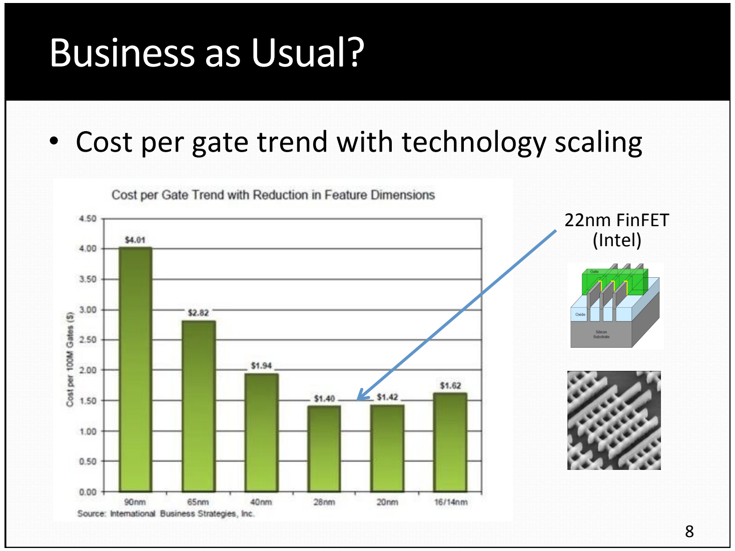
(a) Conservative Scaling

(b) ITRS Scaling

[Esmaeilzadeh et al., ISCA'11]



Multicore and Dark Silicon



Business as Usual?

- Cost per gate trend with technology scaling

Source: International Business Strategies, Inc.

Outline

- Multicore and the power wall
- Energy advantages of hardware accelerators
  - 100-1000x gap in efficiency
  - Do not forget memory!
- Reducing power on adaptive platforms
- Chips go 3D!
- Towards heterogeneous manycores

Energy per operation: 45nm CMOS

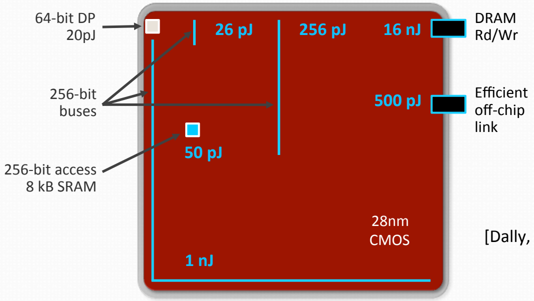
- 32-bit addition: 0.5pJ
- 16-bit multiply: 2.2pJ
- 64-bit FPU: 50pJ/op
- Embedded RISC Processor
  - 32-bit register R/W: 0.33pJ
  - 32-bit cache R/W: 3.5pJ
  - add instruction**: 5.32 pJ

    **add instruction (best case) = fetch, decode, read 2 operands from RF, execute, write back (into local reg. first, then copy into RF)

[Dally et al., Computer, 2010]

The Energy Cost of Data Movement

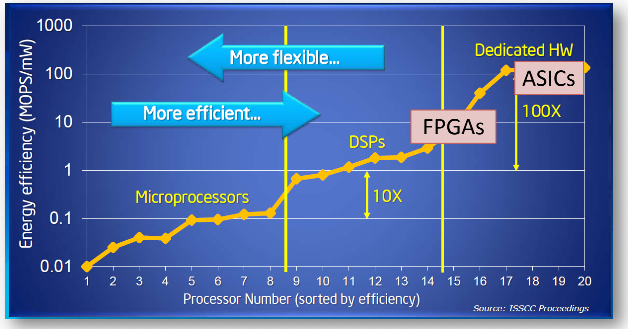
- Fetching operands costs more than computing



[Dally, IPDPS'11]

- Energy cost of cache coherence is huge!

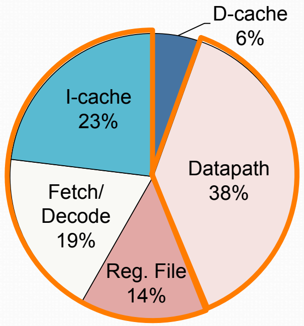
The Efficiency of Specialization



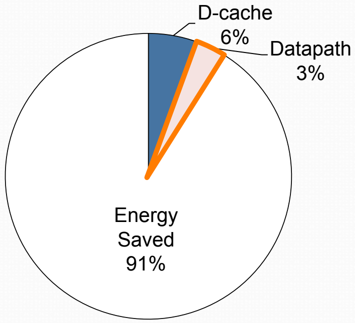
* Source: Ning Zhang and Bob Brodersen, ISSCC data

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

Where do the energy savings come from?



MIPS baseline
91 pJ/instr.

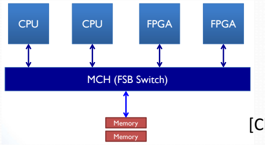


Specialized core
8 pJ/instr.

[Goulding et al., Hot Chips'10]

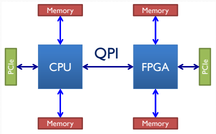
From Embedded Systems to Data Centers

- Many datacenter applications can be accelerated
  - Web search, data mining, database access (e.g. SQL domain-by aggregation)
  - Information security, crypto (e.g. Fully Homomorphic Encryption)
  - Financial, video processing, etc.
- Acceleration in FPGA can keep flexibility while increasing energy efficiency
  - Issue of bandwidth/latency between CPU and FPGA



Intel Front-Side Bus (FSB)

[Chow, FPT'13]



Intel Quick Path Interco. (QPI)

Energy per operation: 40nm V6 FPGA

- 16/32-bit multiply and accumulate:
  - 114pJ (DSP blocks)
  - 170pJ (LUT)
- 32-bit I/O access: 1.47nJ
- 32-bit memory read: 660 pJ
- 32-bit register R/W: 1.12 pJ
- Embedded microblaze processor
  - 16/32-bit multiply and accumulate: 7.4uJ

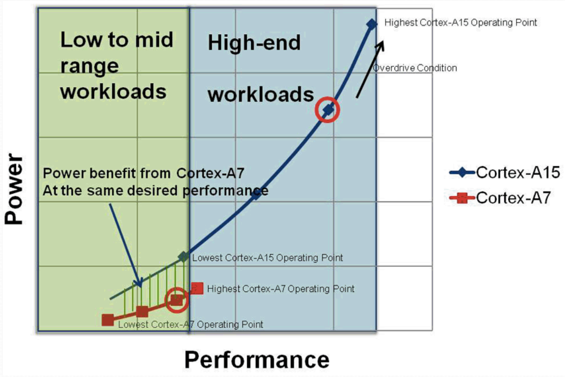
[Bonamy et al., 2013]

Outline

- Multicore and the power wall
- Energy advantages of hardware accelerators
- Reducing power on adaptive platforms
  - Dynamic voltage and frequency scaling
  - Playing with accuracy of operations
  - Sub-word parallelism / SIMD
- Chips go 3D!
- Towards heterogeneous manycores

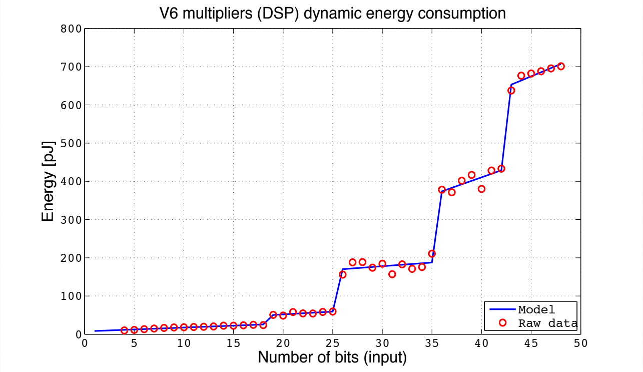
Dynamic Voltage Frequency Scaling

- ARM Big.Little



Energy vs. size in FPGAs (Virtex6, 40nm)

- Multiplier (DSP Blocks)



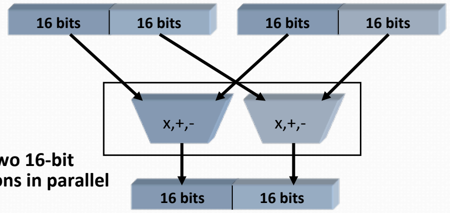
Energy vs. size

- Wire energy
  - 240fJ/bit/mm per transition
  - 32 bits, 10mm: 40pJ/word
  - 8 bits, 10mm: 10pJ/word
- Memory
  - Energy depends on word-length
  - Multiple word access

Subword Parallelism (SWP)*

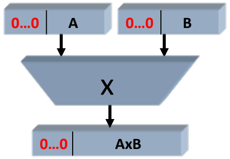
*also called subword-parallel SIMD

- Parallel operations on reduced-precision data
  - Data (sub-words) are packed into words processed by the execution unit in parallel [Fri00]
- Parallel processing increases energy efficiency
  - Trade-off between accuracy and parallelism level

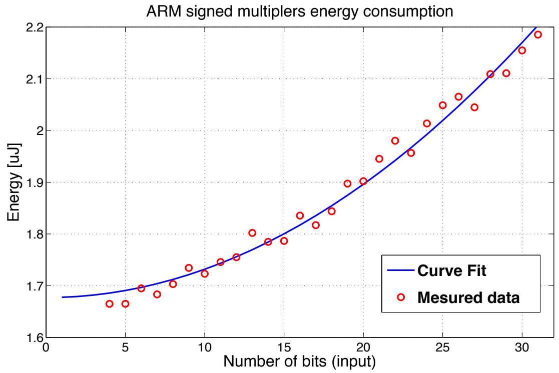


Example: Two 16-bit SWP-operations in parallel

Energy vs. size (ARM)



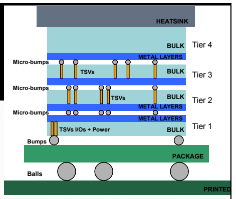
- Signed multiplication



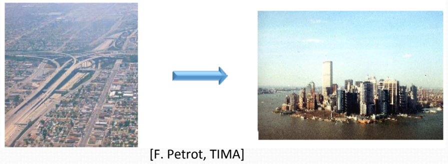
Outline

- Multicore and the power wall
  - Dark Silicon
- Energy advantages of hardware accelerators
  - 100-1000x gap in efficiency
  - Do not forget memory
- Reducing power on adaptive platforms
  - Dynamic voltage and frequency scaling
  - Playing with accuracy of operations
  - Sub-word parallelism / SIMD
- Chips go 3D!
- Towards heterogeneous manycores

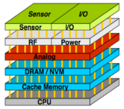
Chips go 3D



- 3D Integrated Circuits
  - Stack Multiple Dies
  - Connect Dies with Through Silicon Vias (TSV)



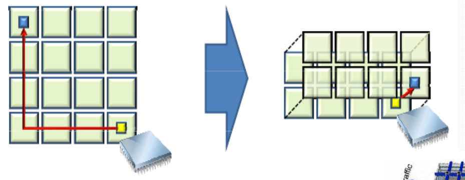
[F. Petrot, TIMA]

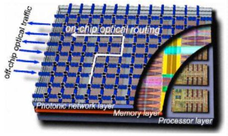


- Examples
  - Image Sensors, Sensor Network Nodes
  - Processor + Memory

Why 3D?

- Wire Length Reduction
  - Replace long, high capacitance wires by TSVs
  - Low latency, low energy, high bandwidth

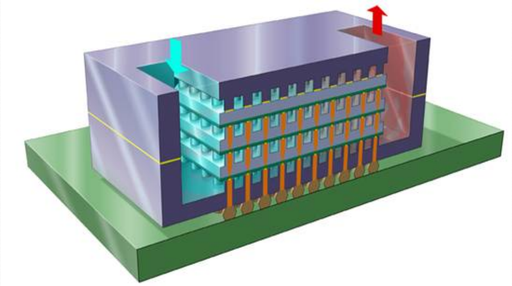




- Small footprint
- Heterogeneous Integration

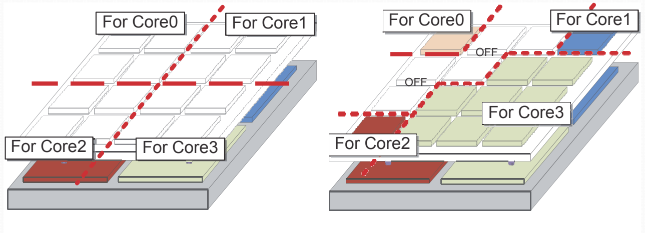
Cooling!

- Thermal effects



3D Memory Stacks

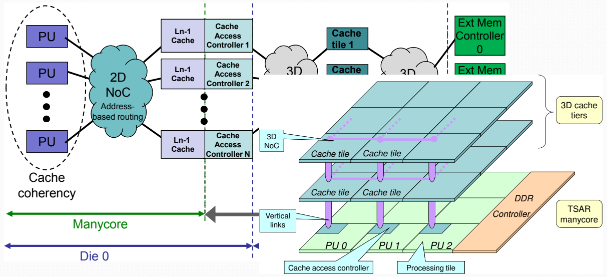
- Moving the compute closer to the data
- Non-Uniform Cache Architecture (NUCA)
  - Dynamic reconfiguration of cache structure



[Jung et al., GLSVLSI'11]

3D Memory Stacks

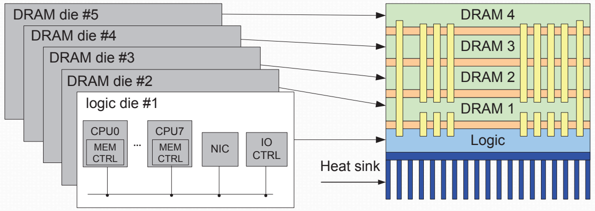
- Moving the compute closer to the data
- Non-Uniform Cache Architecture (NUCA)



[Dutoit et al., DATE'13]

PicoServer

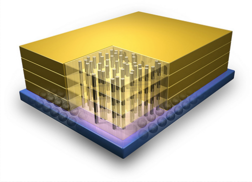
- Energy-efficient multicore architecture with 3D-stacked DRAM

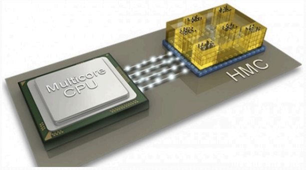


[Kgil et al., SIGOPS'08]

Hybrid Memory Cube

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



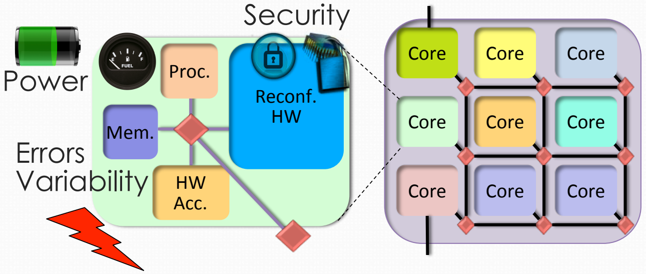


Outline

- Multicore and the power wall
- Energy advantages of hardware accelerators
- Reducing power on adaptive platforms
- Chips go 3D!
- Towards heterogeneous manycores

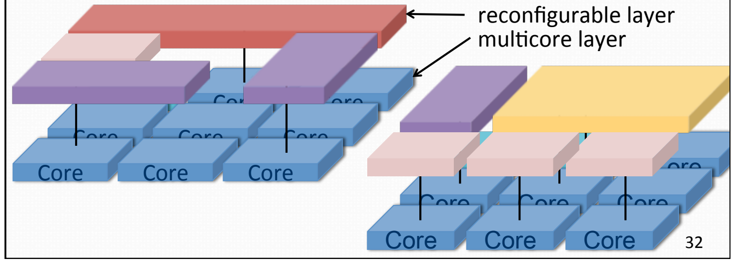
Heterogeneous Multicores

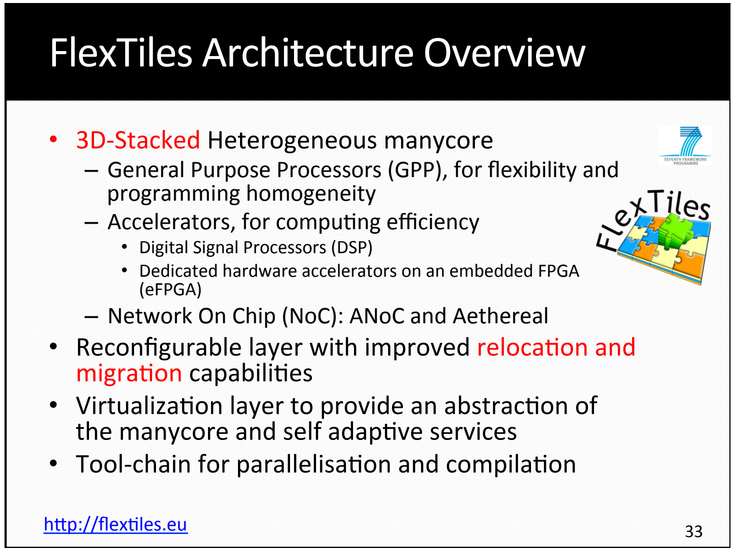
- Different cores on a single chip
  - GPPs, HW accelerators, memory, network-on-chip
- Self-adapting devices
  - Dynamically adapt the hardware to the application
  - Continuously adapt to changing environments



Can 3D Stacking Help?

- 3D-Stacked Reconfigurable Accelerators
  - Improved performance (3D coupling)
  - Improved flexibility
  - Improved resource usage

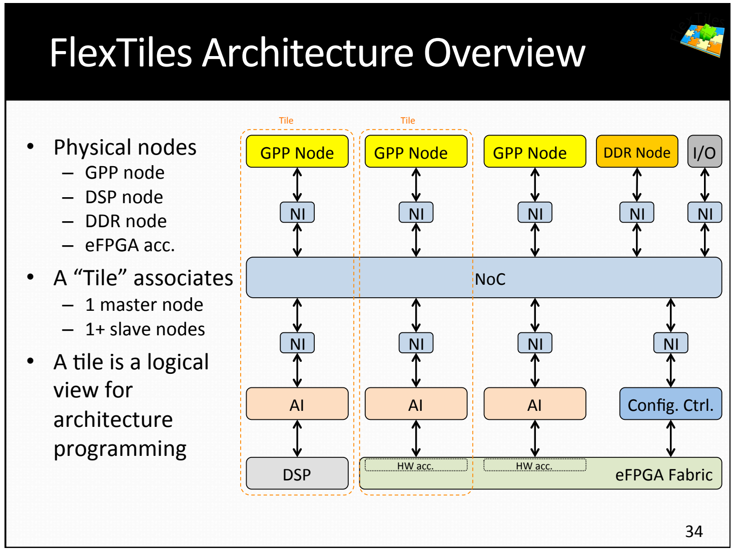




FlexTiles Architecture Overview

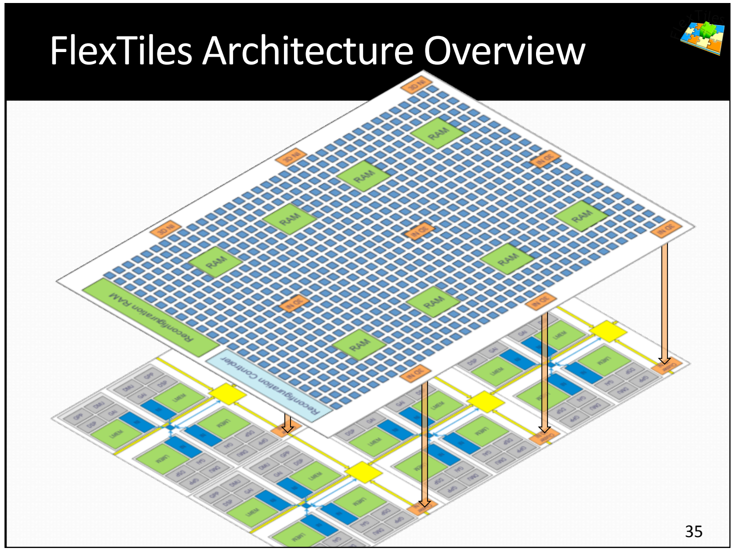
- 3D-Stacked Heterogeneous manycore
  - General Purpose Processors (GPP), for flexibility and programming homogeneity
  - Accelerators, for computing efficiency
    - Digital Signal Processors (DSP)
    - Dedicated hardware accelerators on an embedded FPGA (eFPGA)
  - Network On Chip (NoC): ANoC and Aethereal
- Reconfigurable layer with improved relocation and migration capabilities
- Virtualization layer to provide an abstraction of the manycore and self adaptive services
- Tool-chain for parallelisation and compilation

http://flextiles.eu



FlexTiles Architecture Overview

- Physical nodes
  - GPP node
  - DSP node
  - DDR node
  - eFPGA acc.
- A "Tile" associates
  - 1 master node
  - 1+ slave nodes
- A tile is a logical view for architecture programming



FlexTiles Architecture Overview

Conclusions

- The end of multicore?
  - At least an exciting time for computer architects to deliver performance and efficiency gains
- Dark Silicon for hardware accelerators
- Human Brain
  - 100 trillion synapses @ 20 W!
  - Very "dark" circuits
- Does *The Last Programmer Standing* will be holding an FPGA?

Conclusions

- Bring a new demand for genuinely high level synthesis tools that map programs to circuits
  - Focus on applications rather than compute kernels
  - Able to compile dynamic data structures, recursion and very heterogeneous parallelism
- Domain Specific Languages (DSLs)
  - Can we devise a set of languages to program heterogeneous computing systems?

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.

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