Processors and sockets: What's next?

Greg Astfalk / Salishan Conference / April 25, 2013

Abstract

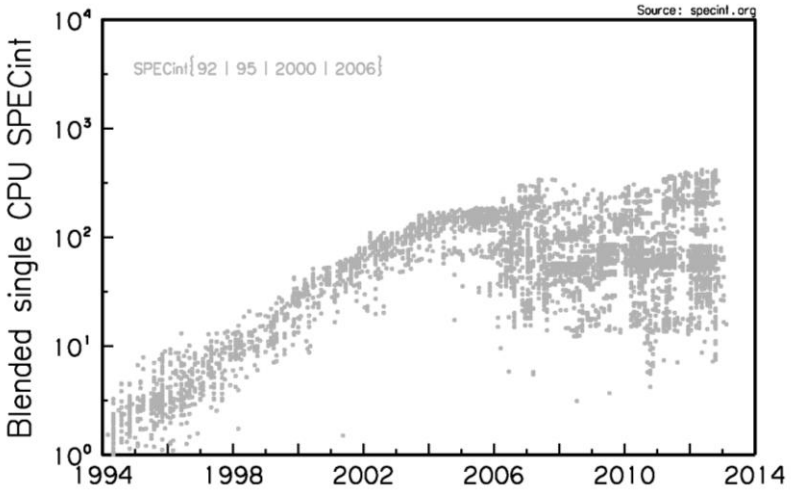
- We have the long-standing notion that, socket equals general-purpose processor. This has served us well, but will it remain the case? There are several maturing technological developments, as well as end-user requirements, which are conspiring to motivate us to change this. The handheld and mobile industry is driving an increasing amount of innovation. Some of this innovation is beginning to exert itself in servers and in the enterprise. We offer a path forward that applies these lower-end technologies and several emerging technologies together. This leads to a picture of what future sockets may become in the next four to five years. We will discuss how they will benefit server applications, and can be relatively easily optimized for different application areas, perhaps even compute-intensive HPC.

Part I

Predicates

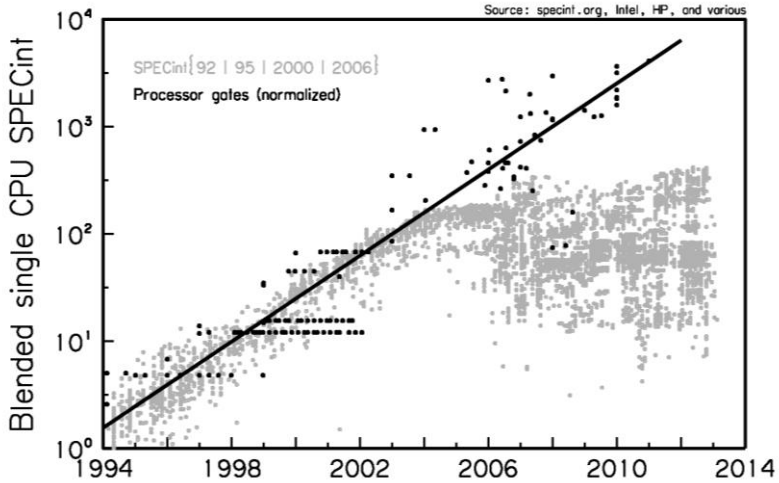
Performance

- Applications cover a broad range
  - Applications and algorithms change and evolve over time
- Accept that there is no universal metric for measuring computer performance
- Let's agree that SPECint is a good proxy
  - Long history so it offers trend information
- We use it as a basis in this talk



Performance, cont'd

- Overlay growth in processor transistors
- Increase in transistors for the past ~9 years has not given commensurate performance benefit
  - >10x increase in gates
  - 10x "drop" in performance
- This is a predicate for this talk



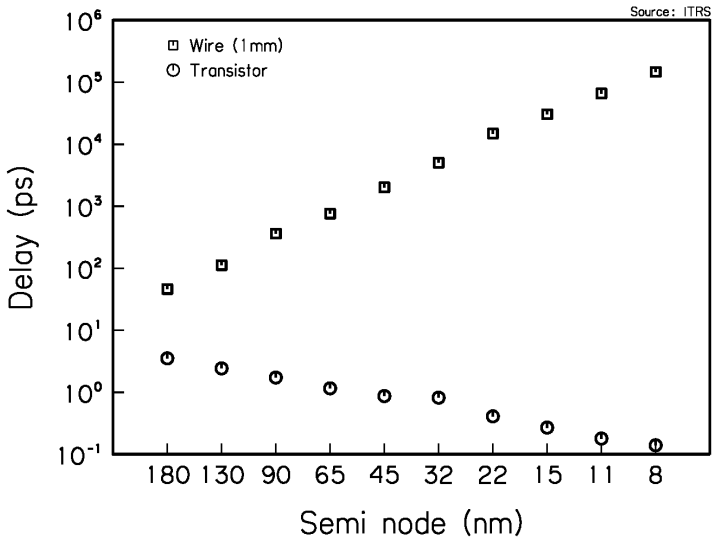
Compute and communication energies

- More energy to move data than to compute on it
  - Computation almost feels “free” relative to communication
  - Time will make this worse
- There are two long poles in the communication energy tent
  - Memory
  - Storage
- This is a predicate for this talk

Operation	Energy (pJ)
64-bit integer operation	1
64-bit floating-point operation	20
256 bit on-die SRAM access	50
256 bit bus transfer (short)	26
256 bit bus transfer (1/2 die)	256
Off-die link (efficient)	500
256 bit bus transfer(across die)	1,000
DRAM read/write	16,000
HDD read/write	$O(10^6)$

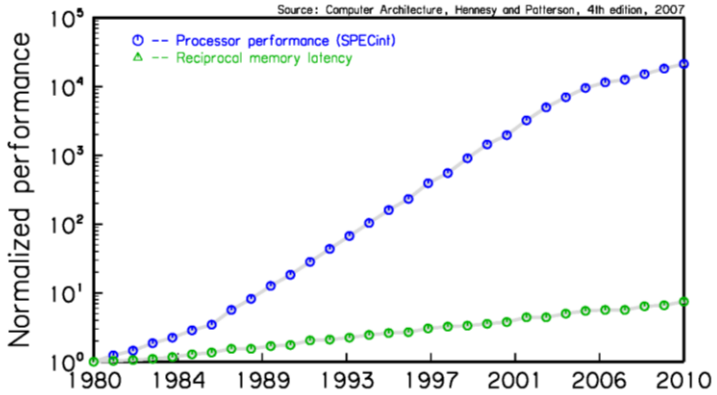
Communication delays

- Transistors continue to get faster
- “Wires” are getting slower
- The gap increases by ~5x with each finer semi node
- This is a predicate for this talk



Memory wall

- Doubtful we need additional evidence of the growing height of the “memory wall”
- This is a significant factor in achieving decent percentages of peak performance on most applications



Observation

- “Those who cannot remember the past are condemned to repeat it“, George Santayana [1905]

- Given the preceding long-standing trends, perhaps it is time to think differently about the issue(s)

Part II

Landscape

ISAs

- This will likely offend somebody…
- The ISA (aka “architecture”) wars are over
  - There are two men left standing
    - x86 and ARM
  - The handful of others are, in aggregate, a minority
- Architecture differences between “CISC” (x86) and “RISC” (ARM) affecting performance are negligible
- There is ample room for innovation and differentiation at the micro-architecture level

The “race to the middle”

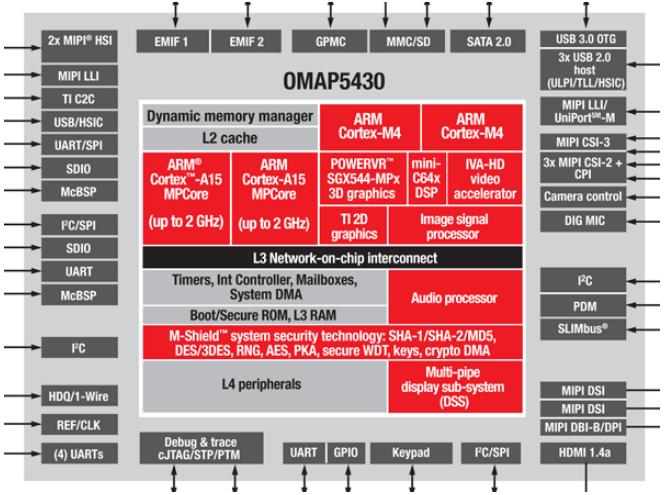
- x86 unarguably owns the high(er) end
- ARM unarguably owns the low(er) end
- Both are moving toward each other in the “middle”
  - x86 attenuating the high-end designs
  - ARM amplifying the low-end designs
- The “middle” is increasing in size and importance
  - The emerging innovation battleground

SoCs

- Yet another industry term subject to misuse and abuse
- We take SoC (System On Chip) to mean
  - "All" the required electronic circuits to build a fully functioning system on a single silicon die
  - There is still latitude in what constitutes "all"
- Is: An ARM-based device in a smart phone
- Is-not: A Xeon with integrated memory controller

SoC IP blocks

- An SoC is a, perhaps large, collection of IP blocks
- If the IP blocks are mature the integration of them is non-trivial, but not challenging
- We can contemplate the integration of new IP blocks as incremental
  - IP block can be <TBD>
  - (Semi-)bespoke SoC



Silicon real estate

- Used McPat do some obvious parameter sweeps

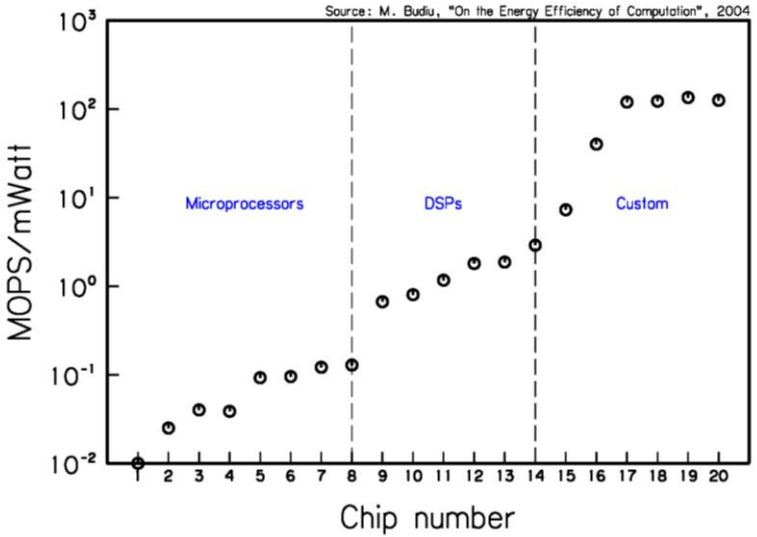
Average percentage of die area					
		Cores	L2	L3	Other
40 nm	GP	47	22	30	1
	SoC	15	38	45	2
32 nm	GP	47	25	27	1
	SoC	15	40	43	2
28 nm	GP	49	24	26	1
	SoC	15	40	43	2

Time constants

- If we contemplate the time to design and manufacture a SoC versus a high-end general purpose processor
  - An SoC can be done in about 1/4 to 1/3 the time
  - The cost, while still low double digit millions, is significantly less
- With an ecosystem of many manufacturers
  - Pipeline multiple designs
  - Multiple concurrent specialized designs

GP vs. specialized vs. custom

- General purpose processors apply 15–20% of their energy to the real work of the algorithm
  - The remaining part is necessary overhead
- An ASIC for a more narrow application space may produce performance benefits of 100–1,000x
  - More area efficient
  - More power efficient
  - This comes at a cost in terms of dollars, time, and product longevity
- Consider a middle-ground solution
  - Semi-specialized SoC



Specialized “sockets”

- Dennard’s scaling has stopped, Moore’s Law continues, enter the “dark silicon” world, thermally capped to add more high-speed gates
- Parallelism is a “one-time gain”
  - Can’t make the cores too small (Amdahl)
  - Can’t lower the voltage too much (threshold)
- Specialization is the only weapon left
  - Integrate more special-purpose functionality
  - Heterogeneous accelerators
- Communication is the real power challenge
  - Integrate more I/O functionality via SoCs

Specialization constraints

- Going toward specialization
  - Applicability is narrower
  - You must shrink the design time
  - Must lower NRE costs
- If you don't
  - You can't really customize effectively
  - Difficult to specialize for something that is 4 years away
    - The market and/or opportunity is gone
  - Benefit to expense ratio goes south

Many-core

- Adding cores helps but incurs a price
  - Performance increases to a point, then declines
    - Ineffectiveness of the caches
    - Limited memory bandwidth per core
  - Caches, dedicated and/or shared, are limited by the transistor count ceiling
  - Memory bandwidth is limited by (pins × frequency)
    - Frequency is capped by power and SI issues
    - Pins are limited and not increasing

Pins

- The number of pins for a socket will not grow significantly
- Assume that 2/3 of the available pins are for power and ground
- We have 1/3 of the pins for supplying off-chip memory bandwidth
- Rather than trying to make more pins or higher frequency pins
  - Avoid the need for pins

Memory

- DRAM will hit EOL (End of Life) in a small number of years
- EOL does not mean you can't buy it next Thursday
  - Capacity scaling plateaus
  - Energy efficiency drops
  - Moves to value-pricing (i.e., you pay more per bit)
  - FIT rate increases
- This will be a long-tailed phenomena

Memory replacement

- DRAM will be replaced by a nonvolatile memory (NVM) technology
- NVM is not like DRAM
  - Latency differences
  - Persistence
  - Write endurance
  - Capacity

DDR

- My belief is that the gorilla in the room is the memory controller
- It is time to think about what comes after DDR4
- DDRx has only three strikes against it
  - Power
  - Performance
  - Pins
- Rethink the memory (aka “DRAM and DDR”) ecosystem

Memory controller IP block

- A new memory protocol and controller as an IP block
- Advances in
  - Packaging
  - Protocol
  - Channel
- will yield
  - Large uplift in memory bandwidth
  - Modest reduction in load-to-use memory latency
  - NVM can give significant increase in memory capacity

On the subject of memory…

- Don’t under-estimate the negative inertia to change a 30+ year old ecosystem
- The volatile to NVM transition and the DDR to DDR++ transition will be slow to unfold
- NVM induces profound changes on many things
  - Memory models
  - Memory consistency
  - Drivers
  - Operating systems
  - Applications
  - Etc.

Part III

Conjectures and suggestions

Conjecture #1

- What seems likely, with high probability, over the next 3–7 years
  - General-purpose sockets will continue to exist
    - With attenuated performance expectations
  - Many-core will emerge and exist
    - (Semi-)heroic programming and/or appropriate applications to exploit the performance
  - SoCs will ascend
    - Better use of gates
    - Greater degree of integration
    - Bespoke designs

Conjecture #2

- What seems likely, with high probability, over the next 4–7 years
  - Memory, as it is practiced today, will change significantly
    - Transition from volatile to nonvolatile
    - Sustainable bandwidth will increase, perhaps by a lot
    - Latency (load-to-use) will decrease modestly
    - Capacity will increase significantly
  - SoCs can be a platform to enable and accelerate this

Suggestions

- Think (channeling Bill Clinton), “It’s the memory, stupid!”
- Think (channeling Bill Clinton), “It’s the communications, stupid!”
- Consider the use of (semi-)purpose-built SoCs
  - Bespoke designs for application benefit
  - Differentiate at the memory controller
    - Better balance (byte/op) → better time-to-solution (conjecture)
    - Wimpier core + strong memory subsystem → better time-to-solution (conjecture)
- Begin to think about and plan for life after DDR and DRAM
- Anything that can be done with algorithms or coding to avoid communication, any communication, is goodness

Acknowledgements

- Far too many people to individually name, HP and elsewhere, have informed me over the years
  - This helped shape my views
- For this talk I want to thank two HP Labs researchers for data
  - Paolo Faraboschi
  - Jichuan Chang

References (partial)

- H. Esmaeilzadeh, etal., Power challenges may end the multicore era, Communications of the ACM , Volume 56 Issue 2, February 2013, pp: 93–102.
- N. Hardavellas, et al., Toward Dark Silicon in Servers, IEEE Micro, Vol.31, No.4, July/August, 2011, pp: 6–15.
- R. Hameed, et al., Understanding Sources of Inefficiency in General-Purpose Chips, Communications of the ACM, Vol.54, No.10, October, 2011, pp: 85–93.
- R. Dennard, J. Cai, and A. Kumar, A perspoective on today's scalaing challenges and possible future directions, Solid-State Electronics, Vol.51, pp: 518–525, 2007.
- Z. Guz, Many-Core vs. Many-Thread Machines: Stay away from the Valley, IEEE Computer Architecture Letter, Vol.8, No.1, pp:25–28, April, 2009.
- B. Grot, D. Hardy, P. Lotfi-Kaman, B. Falsafi, C. Nicopoulos, and Y. Sazeides, Optimizing Data-Center TCO With Scale-Out Processors, IEEE Micro, pp: 52–63, Sep/Oct, 2012.
- E. Blem, J. Menon, and K. Sankaralingam, Power Struggles: Revisiting the RISC vs. CISC Debate on Contemporary ARM and x86 Architectures, HPCA, 2013.
- S. Li, J. Ahn, R. Strong, J. Brockman, D. Tullsen, N. Jouppi, McPAT: An Integrated Power, Area, and Timing Modeling Framework for Multicore and Manycore Architectures, MICRO'09, December 12–16, 2009.

Questions?