Microprocessors -- 10 Years Back, 10 Years Ahead

Guri Sohi

University of Wisconsin — Madison
URL: http://www.cs.wisc.edu/~sohi

Outline

- The enabler: semiconductor technology
- The past 10 years
- The next 10 years
- Wither silicon computing?

The Enabler: Semiconductor Advances

- Shrinkage in feature size
  - more transistors
  - faster transistors
- Increasing die size
  - more transistors

SIA Roadmap

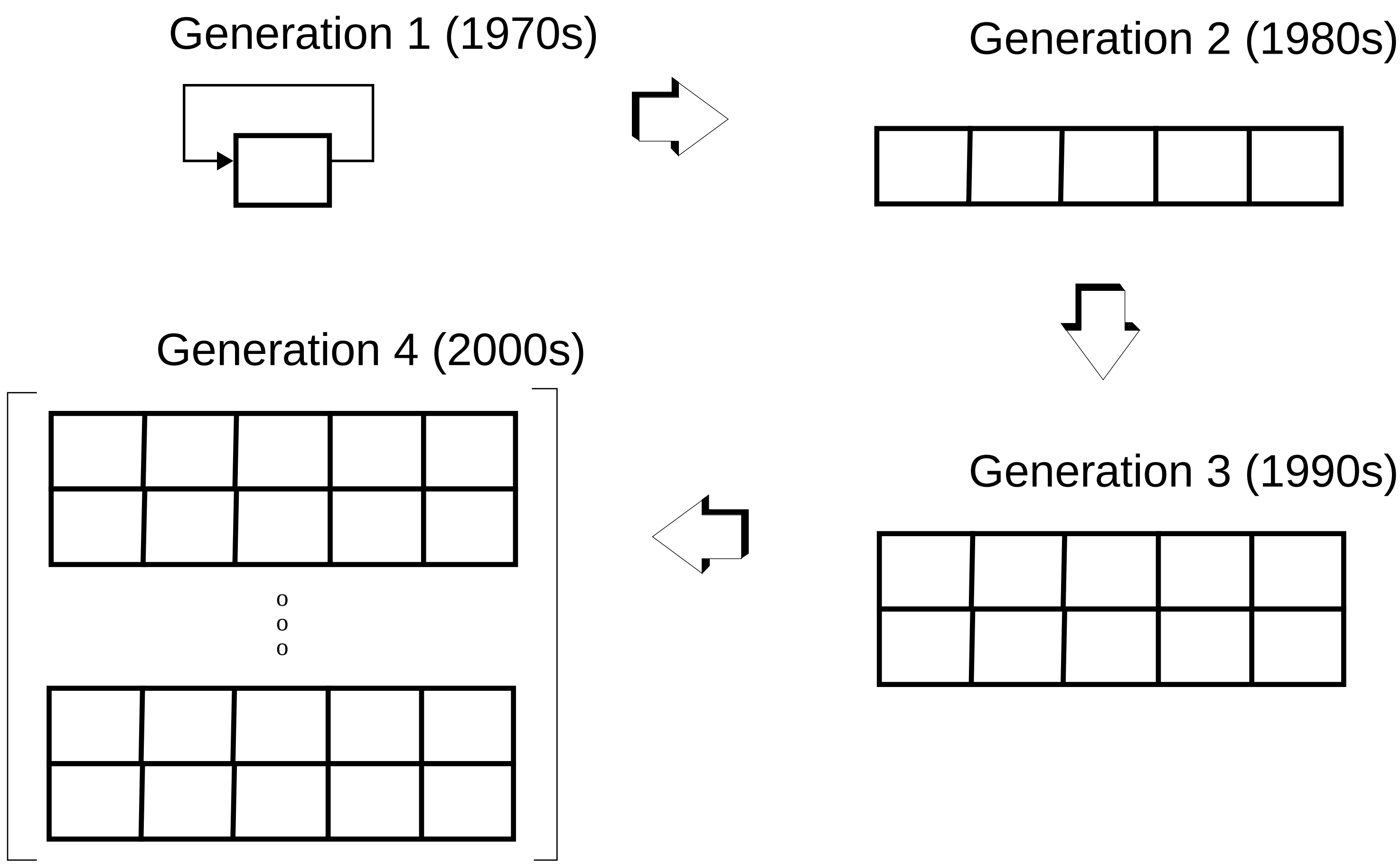
Year	1997	1999	2000	2005	2008	2011	2014
Tech. (nm)	250	180	130	100	70	50	35
Memory(bits)	64M	256M	1G	4G	16G	64G	256G
Logic	3.7M	6.2M	18M	39M	84M	180M	390M

Source: Semiconductor Industry Association (SIA)

Role of Computer Architect

- Use available technology to perform processing tasks
- Match processing tasks to hardware blocks constructed from available technology
- Do so in a manner that is easy to design/verify
- Get desired level of performance

Microprocessor Generations



Microprocessors -- 10 Years Back

- 30X increase in available transistors
  - how to use them?
- Little change in software programming model (still write programs in sequential languages
- Failed promise of automatic parallelization
- Great investment in existing software

Resort to low-level, instruction level parallelism (ILP)

Instruction Level Parallelism

- Determine small number (10-40) instructions to be executed
  - control dependences (branches) hinder determination
- Determine dependence relationships and create dependence graph
- Use dependence graph to execute instructions in parallel
- Can be done statically (VLIW/EPIC) or dynamically (out-of-order (OOO) superscalar

Key: determining which instructions to execute

Use speculation: control speculation

Speculation and Computer Architecture

Speculation: “.. to assume a business risk in hope of gain”

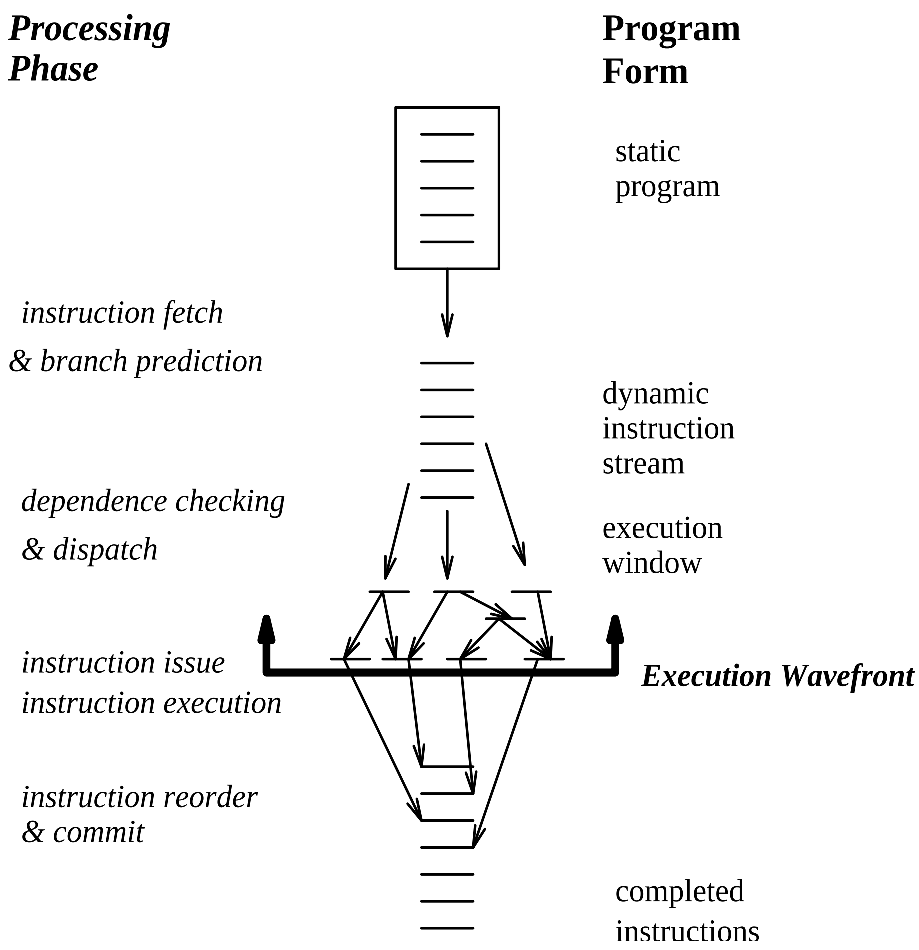
-- Webster

- Speculation in computer architecture is used to try to overcome constraining conditions

Speculation and Computer Architecture

- Speculate outcome of event rather than waiting for outcome to be known
  - mis-speculation if wrong
  - mis-speculation can have penalty
- Develop techniques to speculate better

Model for Out-of-Order Processors



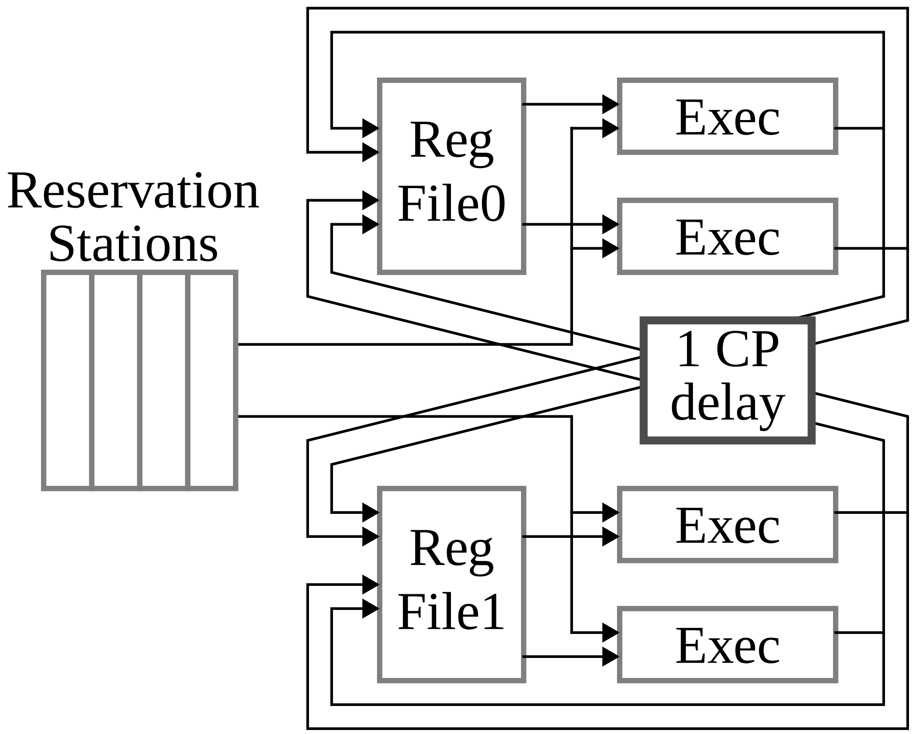
Performance-Inhibiting Constraints

- Brought on by dependences
- Control dependences: inhibit creation of instruction window
  - use control speculation
- Ambiguous data dependences: inhibit parallelism recognition
  - use data dependence speculation
- True data dependences: inhibit parallelism
  - use value speculation

Technology Trends

- Wires used to pass values
- Wires getting relatively slower
  - Short wires for fast clock
  - Short wires implies localized communication

Alpha 21264


Reservation Stations
Reg File0
Reg File1
Exec
Exec
Exec
Exec
1 CP delay

Microprocessors -- the Next 10 Years

- Factor of 30 increase in semiconductor resources
  - how to use it?
- New constraints
  - power consumption
  - wire delays
  - design/verification complexity
- New applications?

Future Processor Architectures

- Engineering considerations will imply computing chips with replicated processing cores
  - a.k.a “multiprocessor” or “multiprocessor-like” or “multithreaded”
- How to assign work to multiple processing cores?
  - independent programs (or threads)
  - parts of a single program

Parallel processing of single program

- Will the promise of explicit/automatic parallelism come true?
- Will new (parallel) programming languages take over the world?

Don't count on it!!!

Speculative Parallelization

- Sequential languages aren't going away
- Use speculation to overcome inhibitors to "automatic" parallelization
- Divide program into "speculatively parallel" portions, or "speculative threads"

Speculative Threads

- Subject of extensive research today
  - different thread types being discovered/investigated
- Several research examples (e.g., Wisconsin Multiscalar, Stanford Hydra)
- Two recent commercial examples
  - Sun Multithreaded Architecture for Java Computing (MAJC) -- circa 1999
  - NEC Merlot -- circa 2000

Generic circa 2010 microprocessor

- 4-8 general-purpose processing engines on chip
  - used to execute independent programs
  - explicitly parallel programs (when possible)
  - speculatively parallel threads
  - helper threads
- Special-purpose processing units (e.g., DSP functionality)
- Elaborate memory hierarchy
- Elaborate inter-chip communication facilities

Circa 2010 microprocessor

- Will run “sequential” program
- Will do so 50-60 times faster than today

Wither Silicon Computing?

- Silicon technology roadmap only clear until about 2015
  - sufficient time for advances?
- Semiconductor technology approaching physical limits
  - can architecture take over after technology scaling?
    - multiple dice systems?
- Role of computing is same
  - computing now means for facilitating communication