ECE/CS 472/572 Computer Architecture: Special Topics Part 1: Terminology

Prof. Lizhong Chen

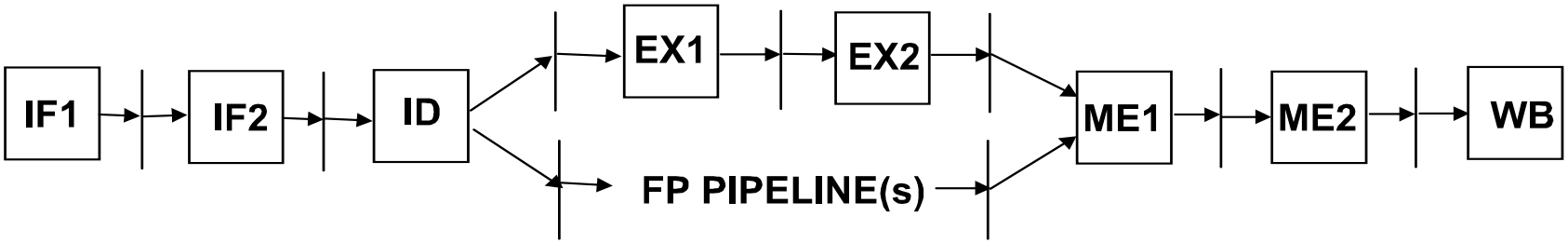
Spring 2019

Level of Parallelism

- Instruction-Level Parallelism (ILP)
  - e.g., pipepline
- Memory-Level Parallelism (MLP)
  - Multiple outstanding cache misses
- Thread-Level Parallelism (TLP)
  - In single core: multi-threading
  - In multiple cores

Superpipeline

- Some stages in the 5-stage integer pipeline are further pipelined
  - To increase the clock rate
  - Not “free”
    - Branch penalty is now 3 clock cycles
    - Instruction latency in cycles is higher



Superscalar

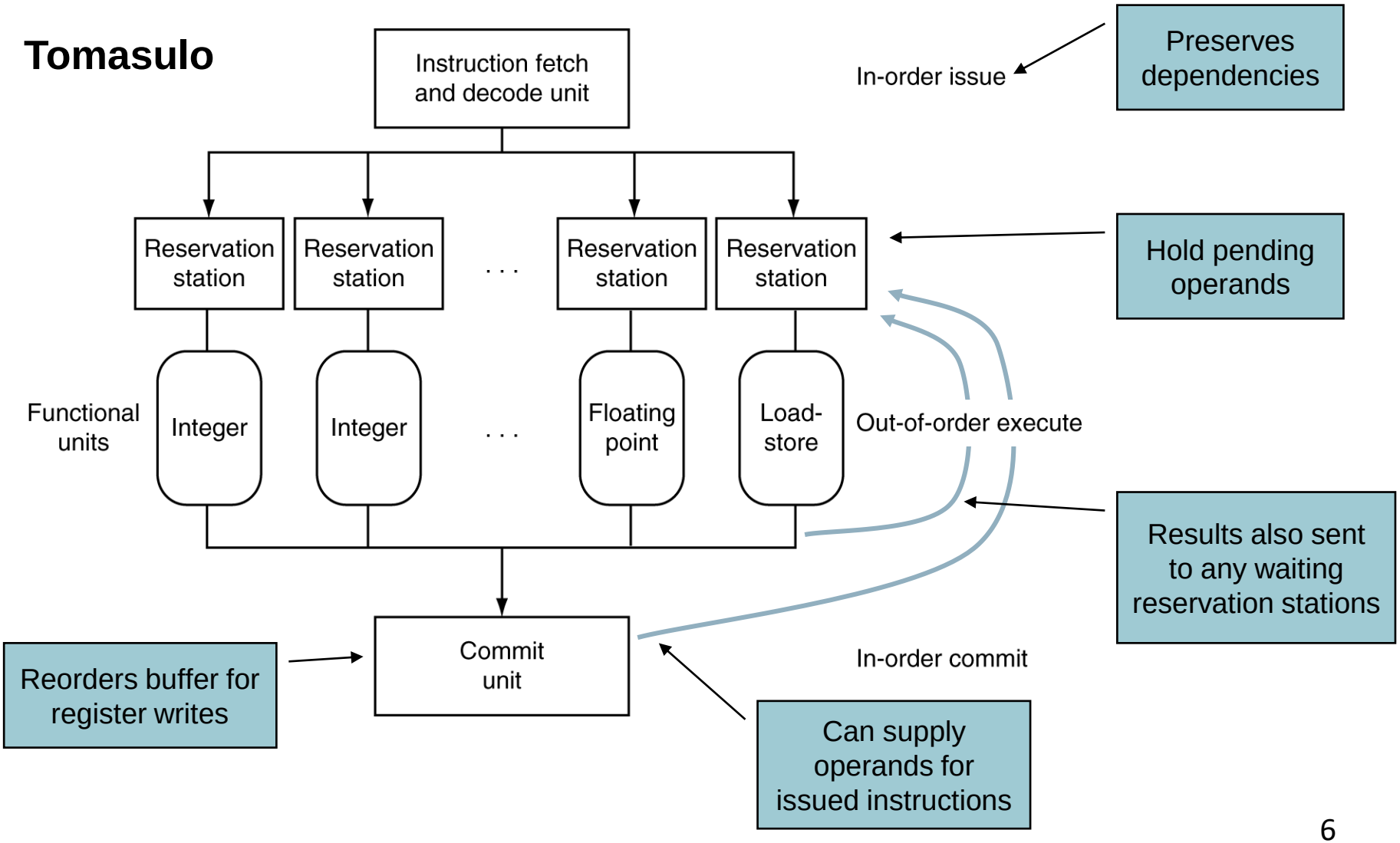
- Also called “multiple issue”
  - Issue multiple instructions in the same cycle
- Static multiple issue
- Dynamic multiple issue

Static Multiple Issue

- 2-way multiple-issue
- Compiler detects and avoids hazards

Address	Instruction type	Pipeline Stages						
n	ALU/branch	IF	ID	EX	MEM	WB		
n + 4	Load/store	IF	ID	EX	MEM	WB		
n + 8	ALU/branch		IF	ID	EX	MEM	WB	
n + 12	Load/store		IF	ID	EX	MEM	WB	
n + 16	ALU/branch			IF	ID	EX	MEM	WB
n + 20	Load/store			IF	ID	EX	MEM	WB

Dynamic Multiple Issue



Speculation

- "Guess" what to do with an instruction
  - Start operation as soon as possible
  - Check whether guess was right
    - If so, complete the operation
    - If not, roll-back and do the right thing
- Examples
  - Speculate on branch outcome
    - Roll back if path taken is different
  - Speculate on load
    - Roll back if location is updated
- Value-prediction

ITRS

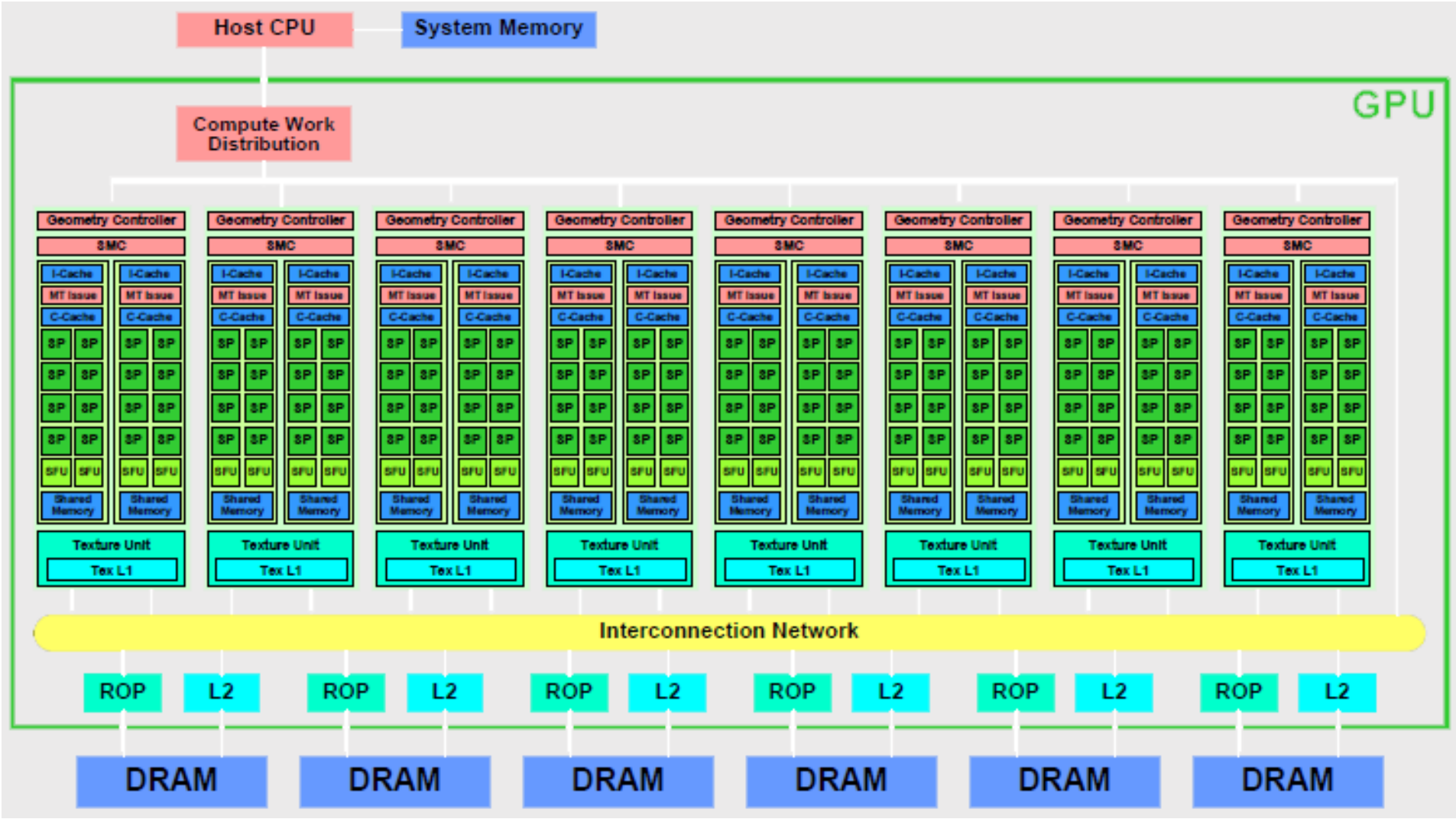
- International Technology Roadmap for Semiconductors

2013	2015	2017	2019	2021	2023	2025	2028
"16/14"	"10"	"7"	"5"	"3.5"	"2.5"	"1.8"	
DRAM ½ Pitch (nm)	28	24	20	17	14	12	10	7.7
FinFET Fin Half-pitch (new) (nm)	30	24	19	15	12	9.5	7.5	5.3
FinFET Fin Width (new) (nm)	7.6	7.2	6.8	6.4	6.1	5.7	5.4	5.0
6-t SRAM Cell Size(um2) [@60f2]	0.096	0.061	0.038	0.024	0.015	0.010	0.0060	0.0030
MPU/ASIC HighPerf 4t NAND Gate Size(um2)	0.248	0.157	0.099	0.062	0.039	0.025	0.018	0.009
4-input NAND Gate Density (Kgates/mm) [@155f2]	4.03E+03	6.37E+03	1.01E+04	1.61E+04	2.55E+04	4.05E+04	6.42E+04	1.28E+05
Flash Generations Label (bits per chip) (SLC/MLC)	64G /128G	128G /256G	256G / 512G	512G / 1T	512G / 1T	1T / 2T	2T / 4T	4T / 8T
Flash 3D Number of Layer targets (at relaxed Poly half pitch)	16-32	16-32	16-32	32-64	48-96	64-128	96-192	192-384
Flash 3D Layer half-pitch targets (nm)	64nm	54nm	45nm	30nm	28nm	27nm	25nm	22nm
DRAM Generations Label (bits per chip)	4G	8G	8G	16G	32G	32G	32G	32G
450mm Production High Volume Manufacturing Begins (100Kwspm)				2018				
Vdd (High Performance, high Vdd transistors)[**]	0.86	0.83	0.80	0.77	0.74	0.71	0.68	0.64
1/(CV/I) (1/psec) [**]	1.13	1.53	1.75	1.97	2.10	2.29	2.52	3.17
On-chip local clock MPU HP [at 4% CAGR]	5.50	5.95	6.44	6.96	7.53	8.14	8.8	9.9
Maximum number wiring levels [unchanged	13	13	14	14	15	15	16	17
MPU High-Performance (HP) Printed Gate Length (GLpr) (nm) [**]	28	22	18	14	11	9	7	5
MPU High-Performance Physical Gate Length (GLph) (nm) [**]	20	17	14	12	10	8	7	5
ASIC/Low Standby Power (LP) Physical Gate Length (nm) (GLph)[**]	23	19	16	13	11	9	8	6

SISD, MIMD, SPMD, SIMD

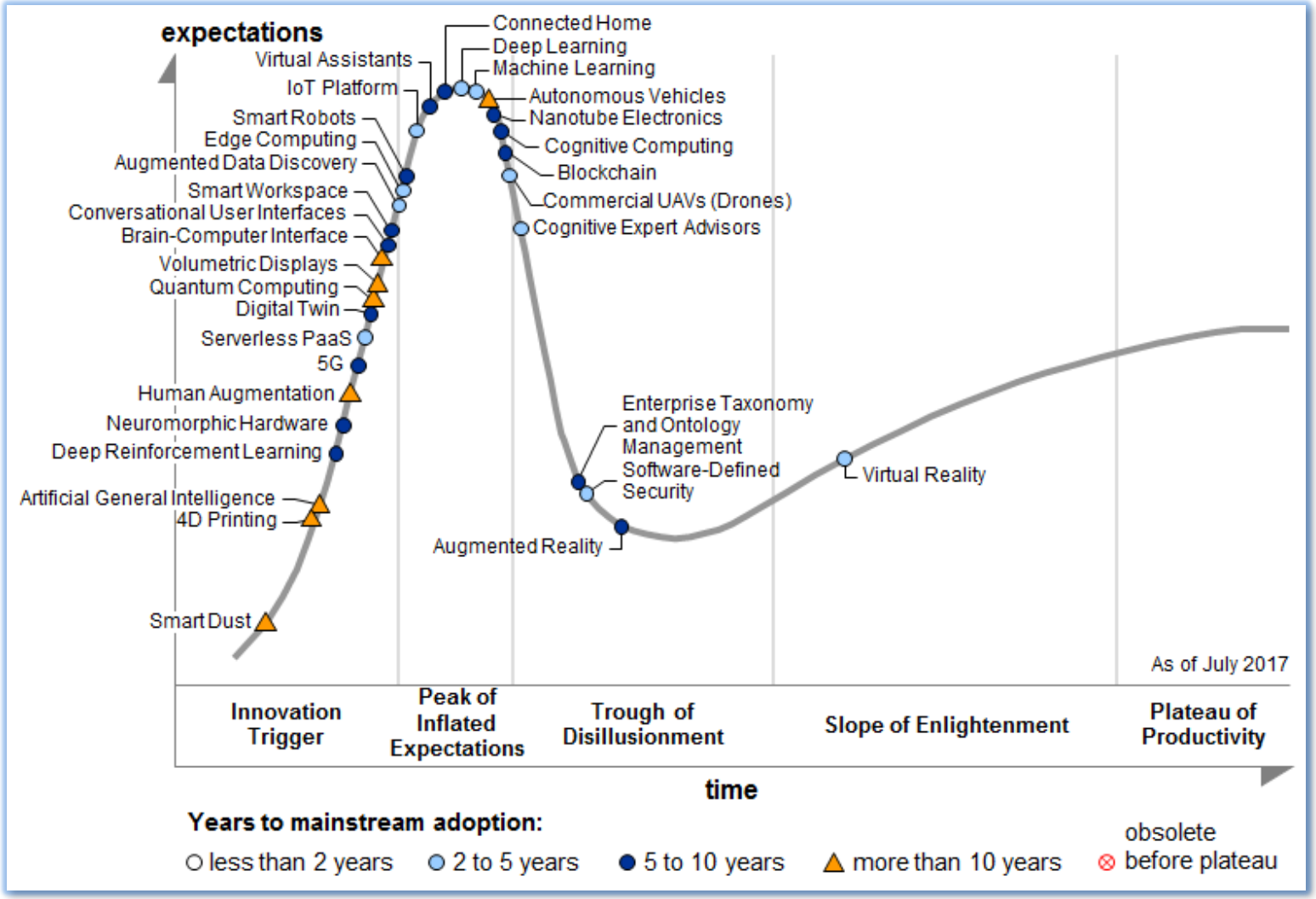
- SISD
  - Single Instruction stream Single Data stream
- MIMD
  - Multiple Instruction streams Multiple Data streams
- SPMD
  - Single Program Multiple Data streams
- SIMD
  - Single Instruction stream Multiple Data streams

GPU Architecture: Nvidia Tesla



The Hype Cycle

"The Hype Cycle offers a glimpse of the future"



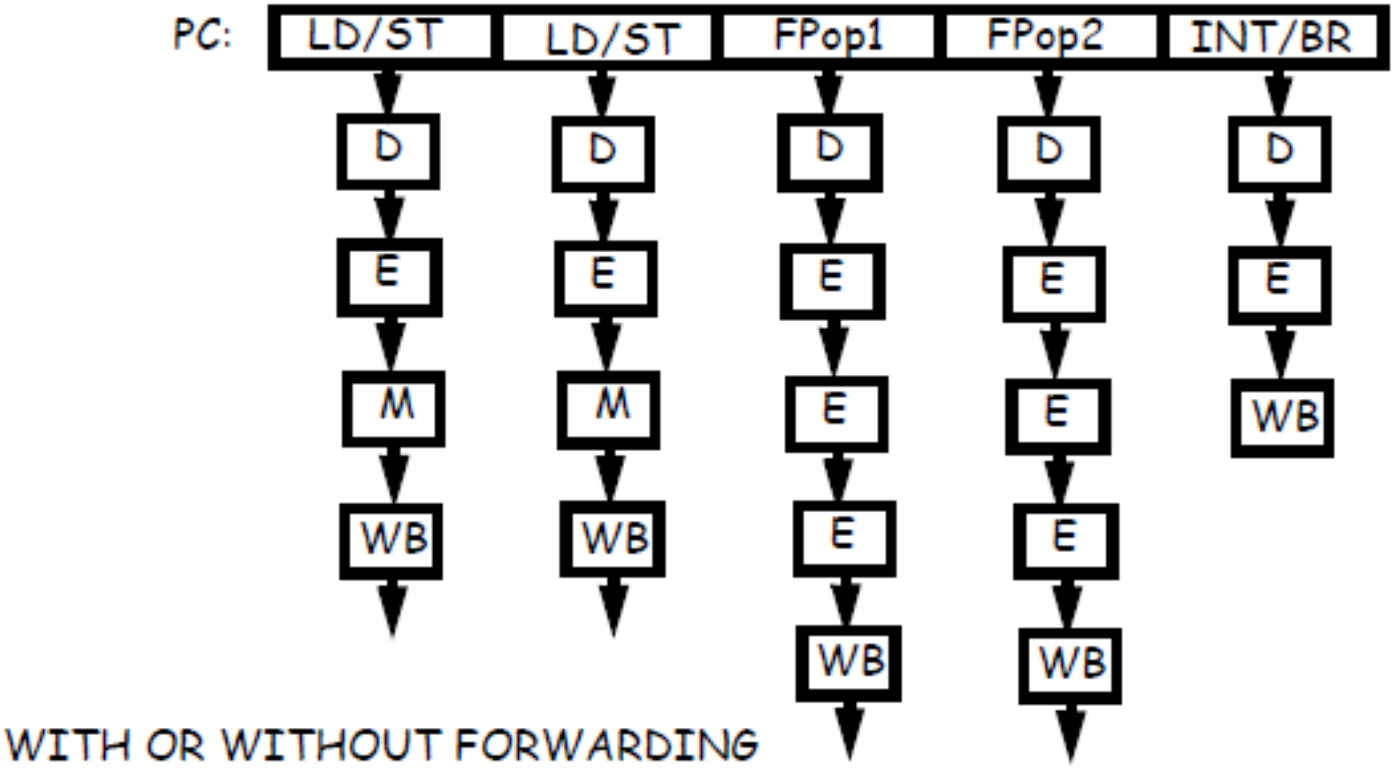
Cortex A8 and Intel i7

Processor	ARM A8	Intel Core i7 920
Market	Personal Mobile Device	Server, cloud
Thermal design power	2 Watts	130 Watts
Clock rate	1 GHz	2.66 GHz
Cores/Chip	1	4
Floating point?	No	Yes
Multiple issue?	Dynamic	Dynamic
Peak instructions/clock cycle	2	4
Pipeline stages	14	14
Pipeline schedule	Static in-order	Dynamic out-of-order with speculation
Branch prediction	2-level	2-level
1st level caches/core	32 KiB I, 32 KiB D	32 KiB I, 32 KiB D
2nd level caches/core	128-1024 KiB	256 KiB
3rd level caches (shared)	-	2- 8 MB

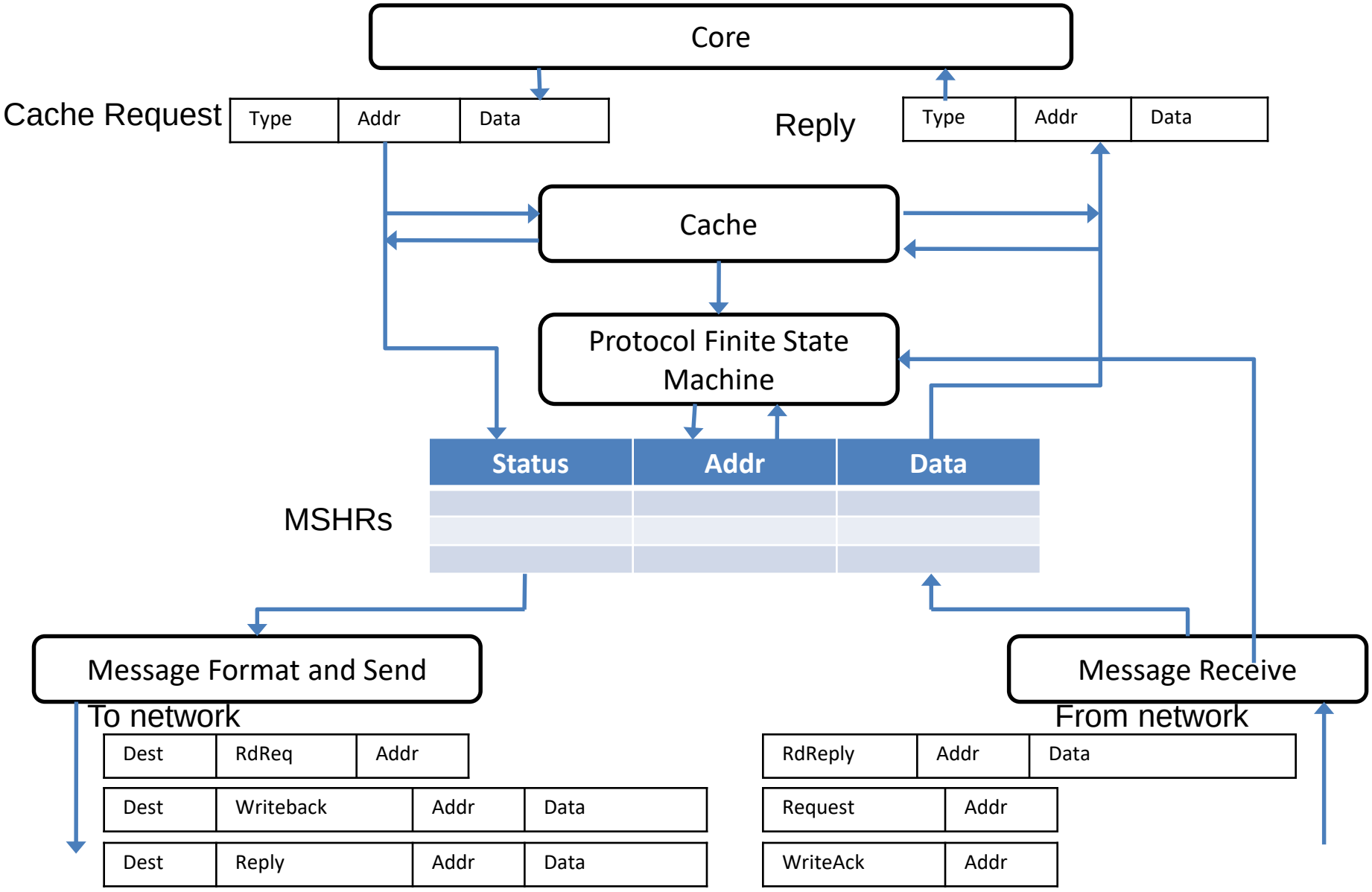
LIW/VLIW

- Long Instruction Word / Very Long Instr. Word
- Multiple independent RISC instructions are packaged in one LIW or VLIW instruction
- Independent functional units
- May have some forwarding to reduce latency

LIW/VLIW


PC:
LD/ST
LD/ST
FPop1
FPop2
INT/BR
D
E
M
WB
WITH OR WITHOUT FORWARDING

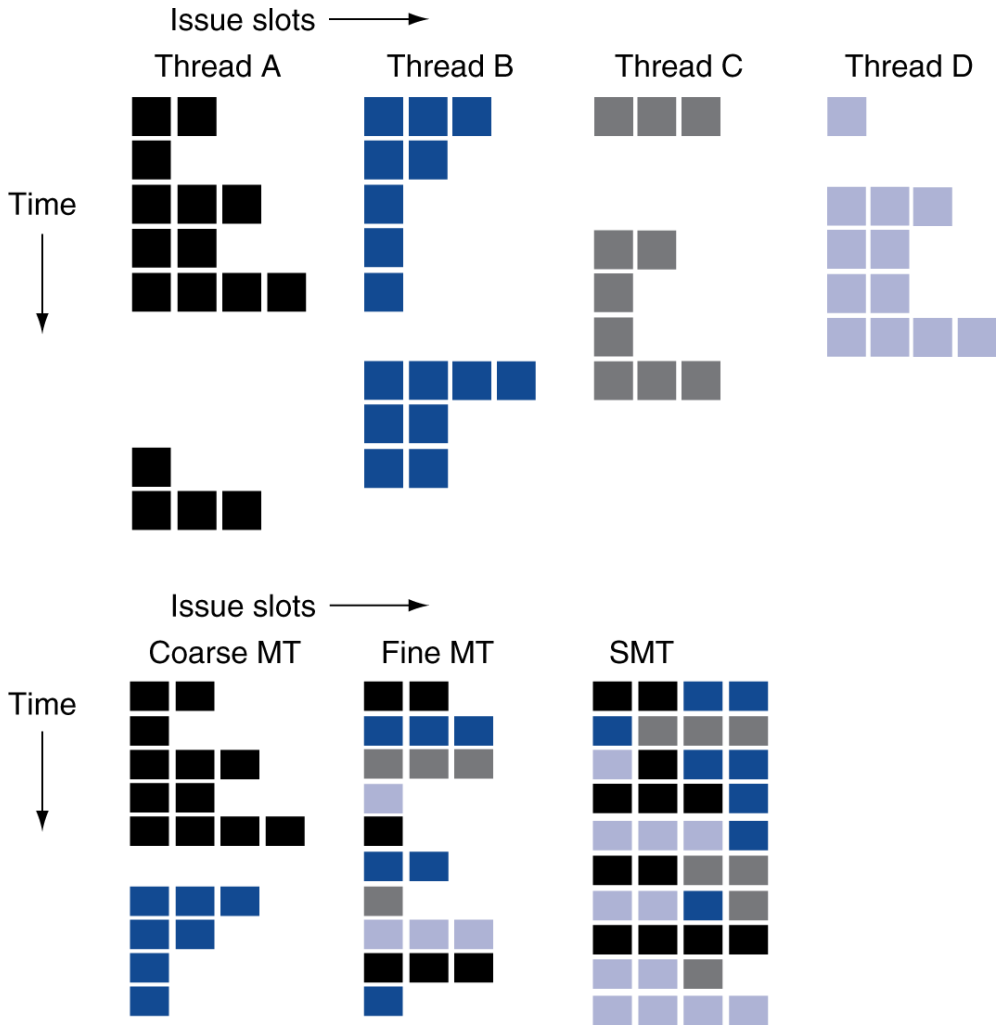
Miss Status Handling Registers (MSHR)



Multithreading

- Fine-grain multithreading
  - Interleave instruction execution
  - If one thread stalls, others are executed
- Coarse-grain multithreading
  - Only switch on long stall (e.g., L2-cache miss)
  - Simplifies hardware, but doesn't hide short stalls (eg, data hazards)
- Simultaneous Multithreading (SMT)
  - In multiple-issue dynamically scheduled processor

Multithreading



Quantum Computers

- "Qubit"
  - A qubit (quantum bit) is the basic unit of quantum information
  - Superposition of 0 and 1
  - Store two bits of information
- Basics of quantum computers
  - https://www.youtube.com/watch?v=JhHMJCUmq28