The Abdus Salam International Centre for Theoretical Physics

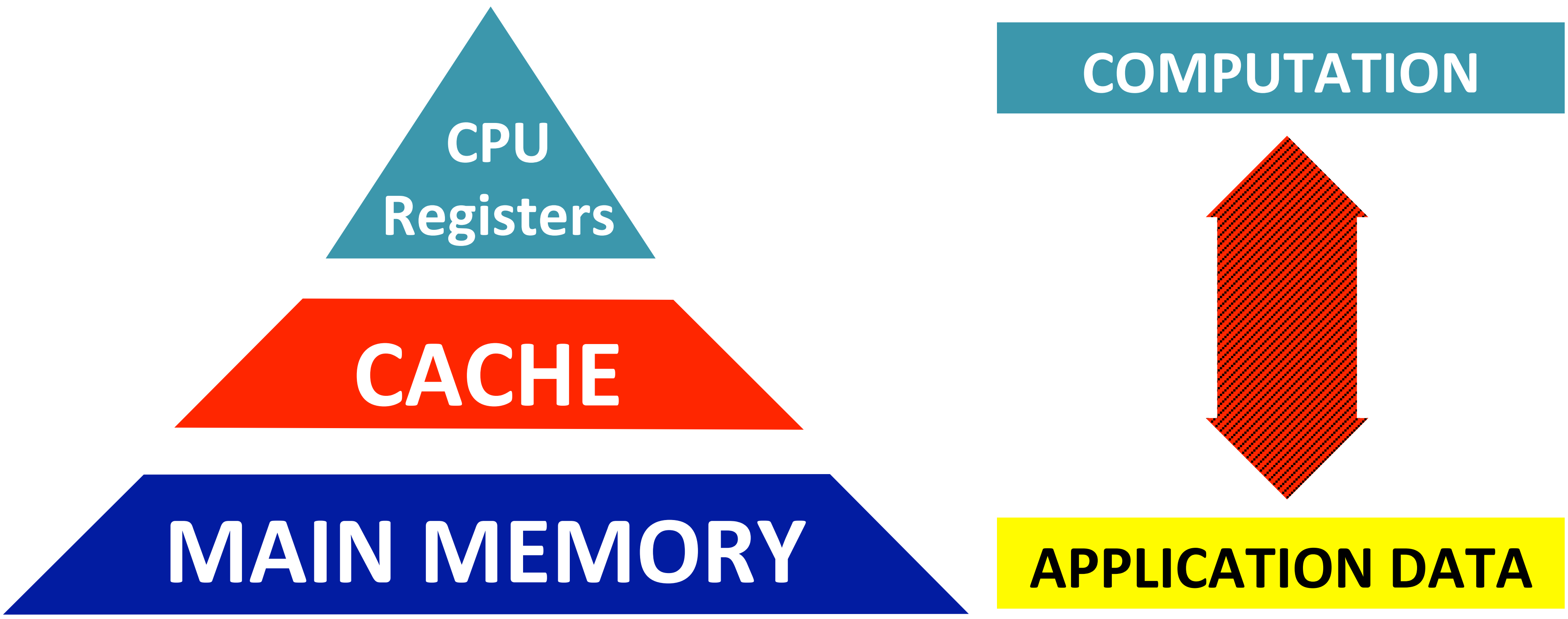
Overview on modern computer architectures: the software crisis

Ivan Girotto – igirotto@ictp.it
Information & Communication Technology Section (ICTS)
International Centre for Theoretical Physics (ICTP)

How fast is my CPU?!

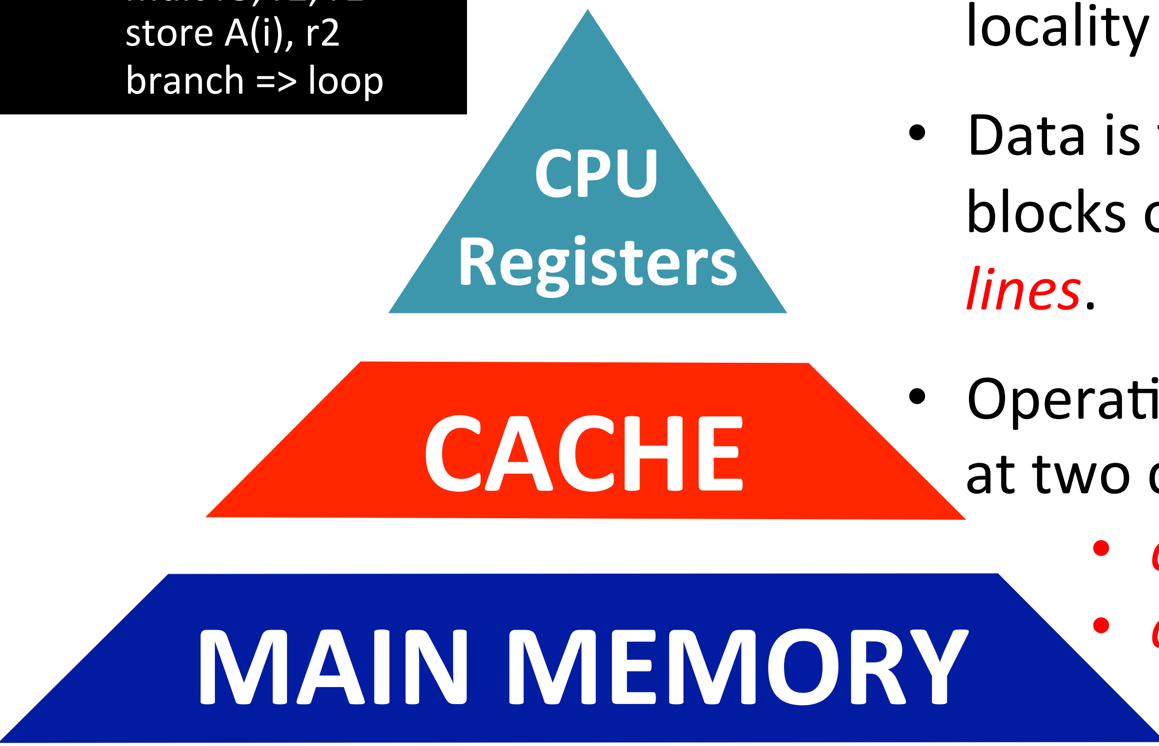
- CPU power is measured in FLOPS
  - number of floating point operations x second
  - FLOPS = #cores x clock x $\frac{\text{FLOP}}{\text{cycle}}$
- FLOP/cycle is the number of multiply-add (FMA) performed per cycle
  - architectural limit
  - depend also by the instruction set supported

The CPU Memory Hierarchy



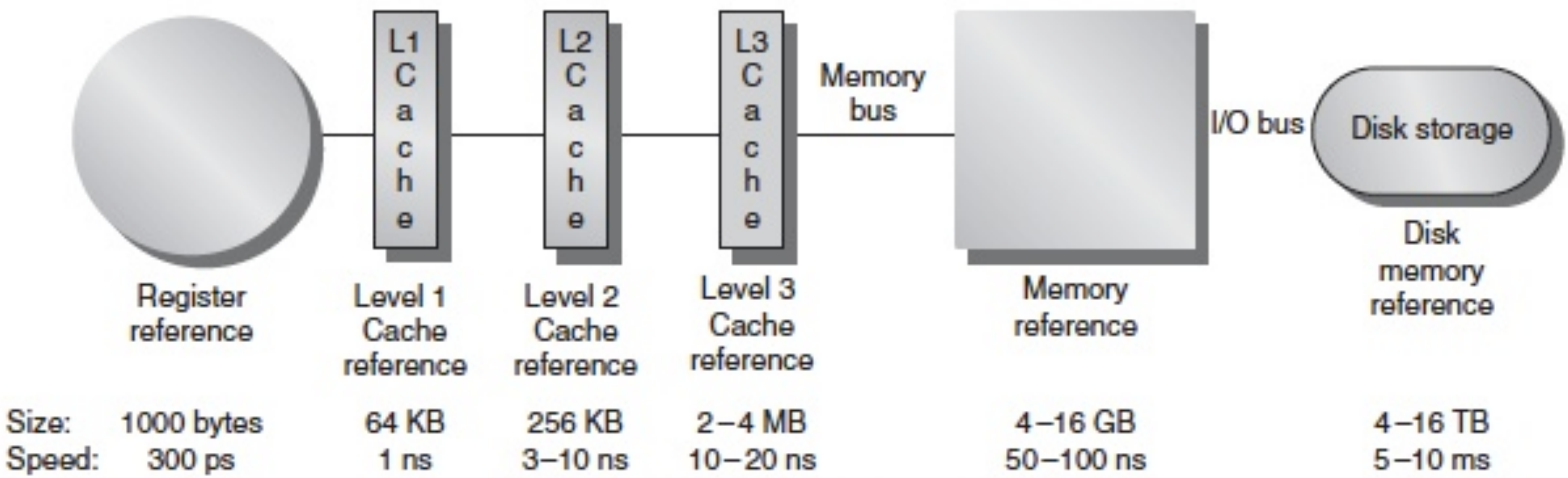
Cache Memory

```
Loop: load r1, A(i)
      load r2, s
      mult r3, r2, r1
      store A(i), r2
      branch => loop
```



- Designed for temporal/spatial locality
- Data is transferred to cache in blocks of fixed size, called *cache lines*.
- Operation of LOAD/STORE can lead at two different scenario:
  - *cache hit*
  - *cache miss*

The CPU Memory Hierarchy



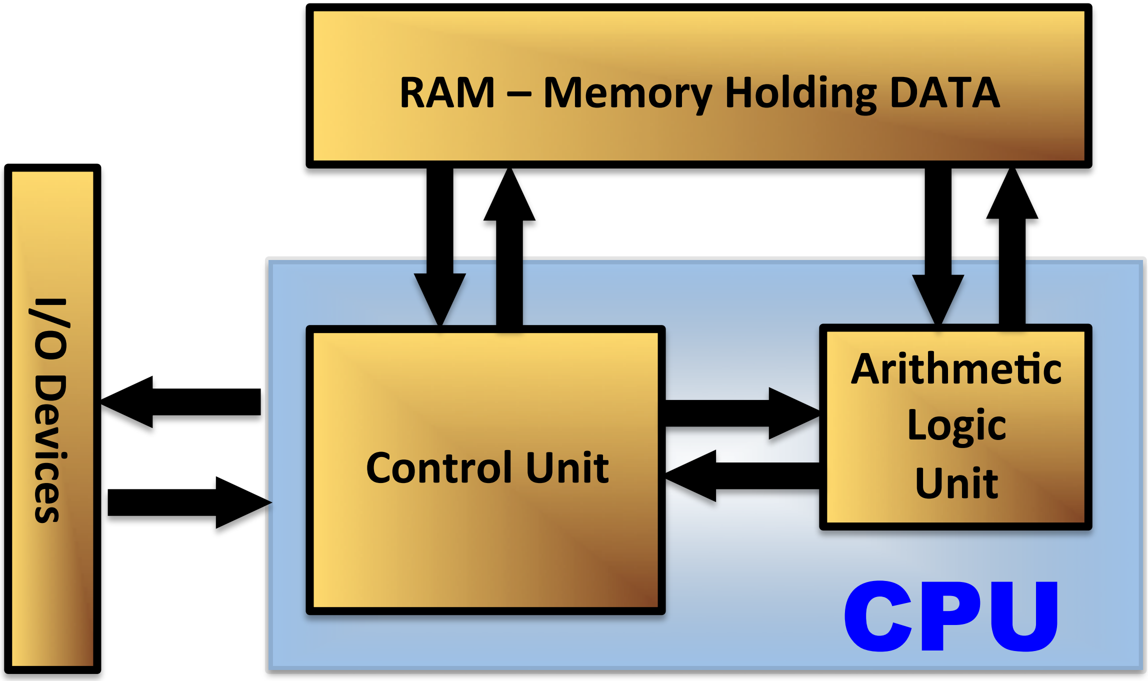
(a) Memory hierarchy for server

Data Memory Access

- Data ordering
- Reduce at minimum the data transfers
- Avoid complex data structure within computational intensive kernels
- Define constants and help the compiler
- Exploit the memory hierarchy

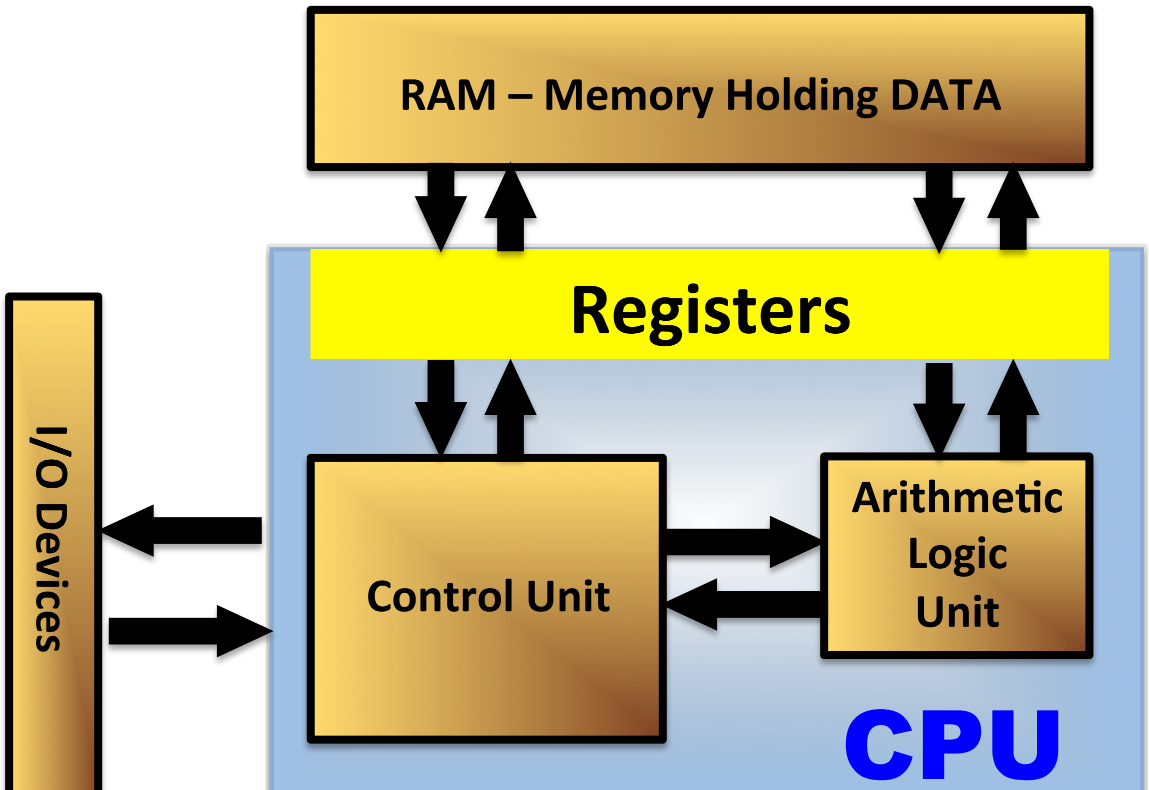
John Von Neumann

The Classical Model

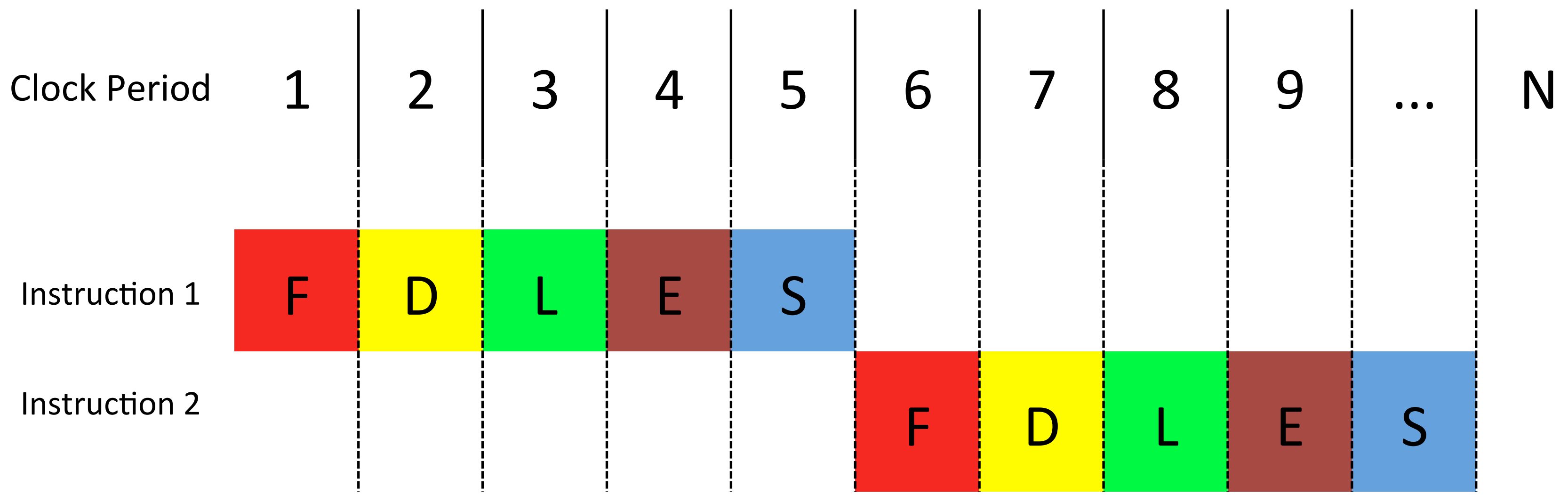


John Von Neumann

The Classical Model



Sequential Processing



Pipelining

Clock Period	1	2	3	4	5	6	7	8	9	...	N
Instruction 1	F	D	L	E	S						
Instruction 2		F	D	L	E	S					
Instruction 3			F	D	L	E	S				
Instruction 4				F	D	L	E	S			
Instruction 5					F	D	L	E	S		

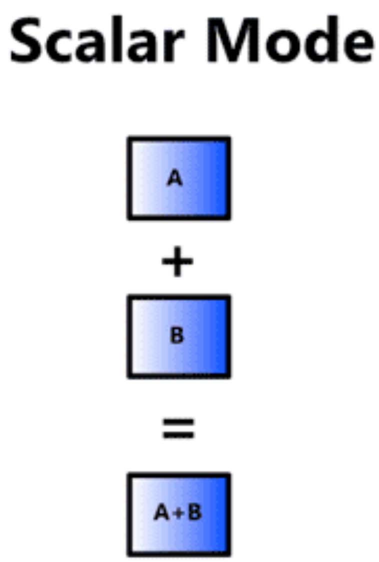
Pipelining

Clock Period	1	2	3	4	5	6	7	8	9	...	N
Instruction 1	F	D	L	E	S						
Instruction 2		F	D	L	E	S					
Instruction 3			F	D	L	E	S				
Instruction 4				F	D	L	E	S			
Instruction 5					F	D	L	E	S		

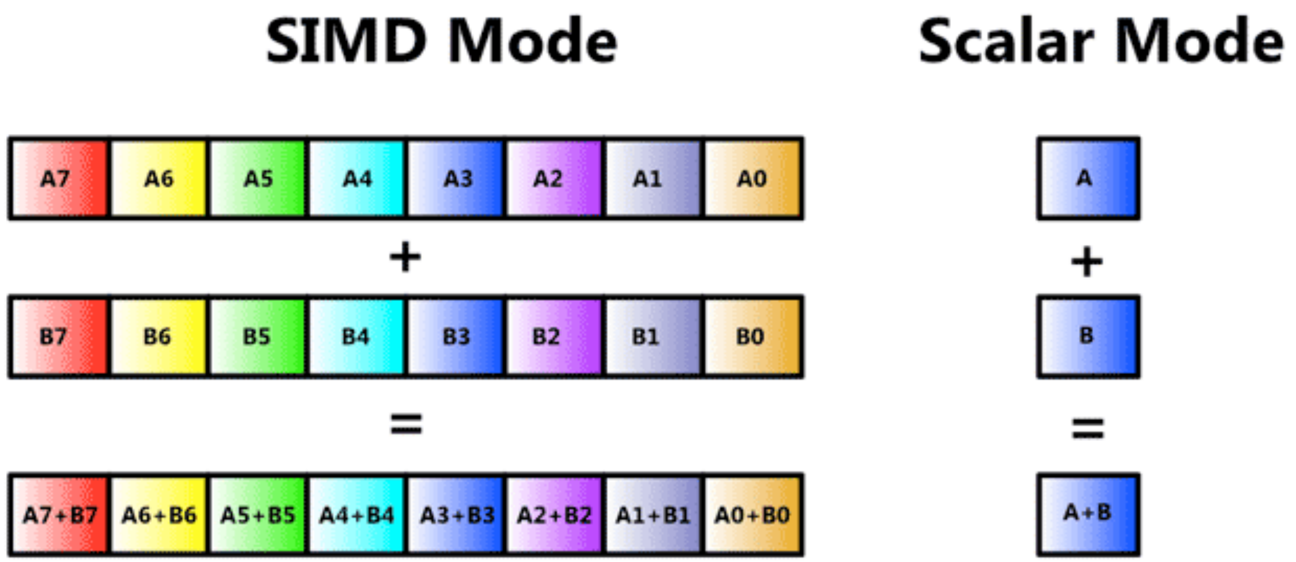
Superscalaring

Clock Period	1	2	3	4	5	6	7	8	9	...	N
Instruction 1	F	D	L	E	S						
Instruction 2	F	D	L	E	S						
Instruction 3		F	D	L	E	S					
Instruction 4		F	D	L	E	S					
Instruction 5			F	D	L	E	S				
Instruction 6			F	D	L	E	S				

The Inside Parallelism



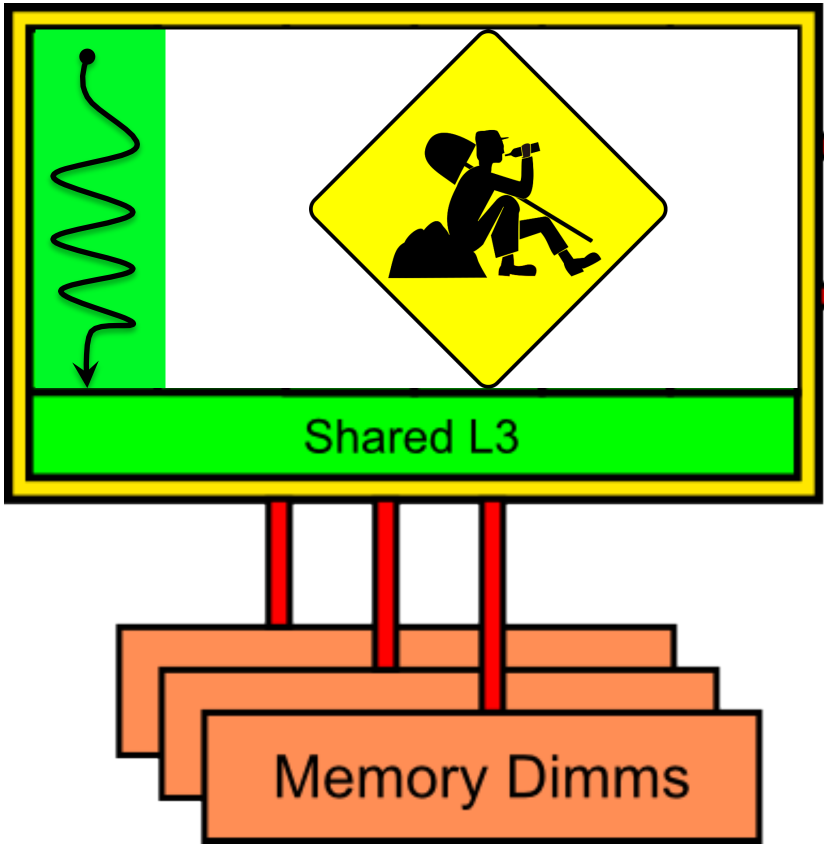
The Inside Parallelism



Performance Metrics

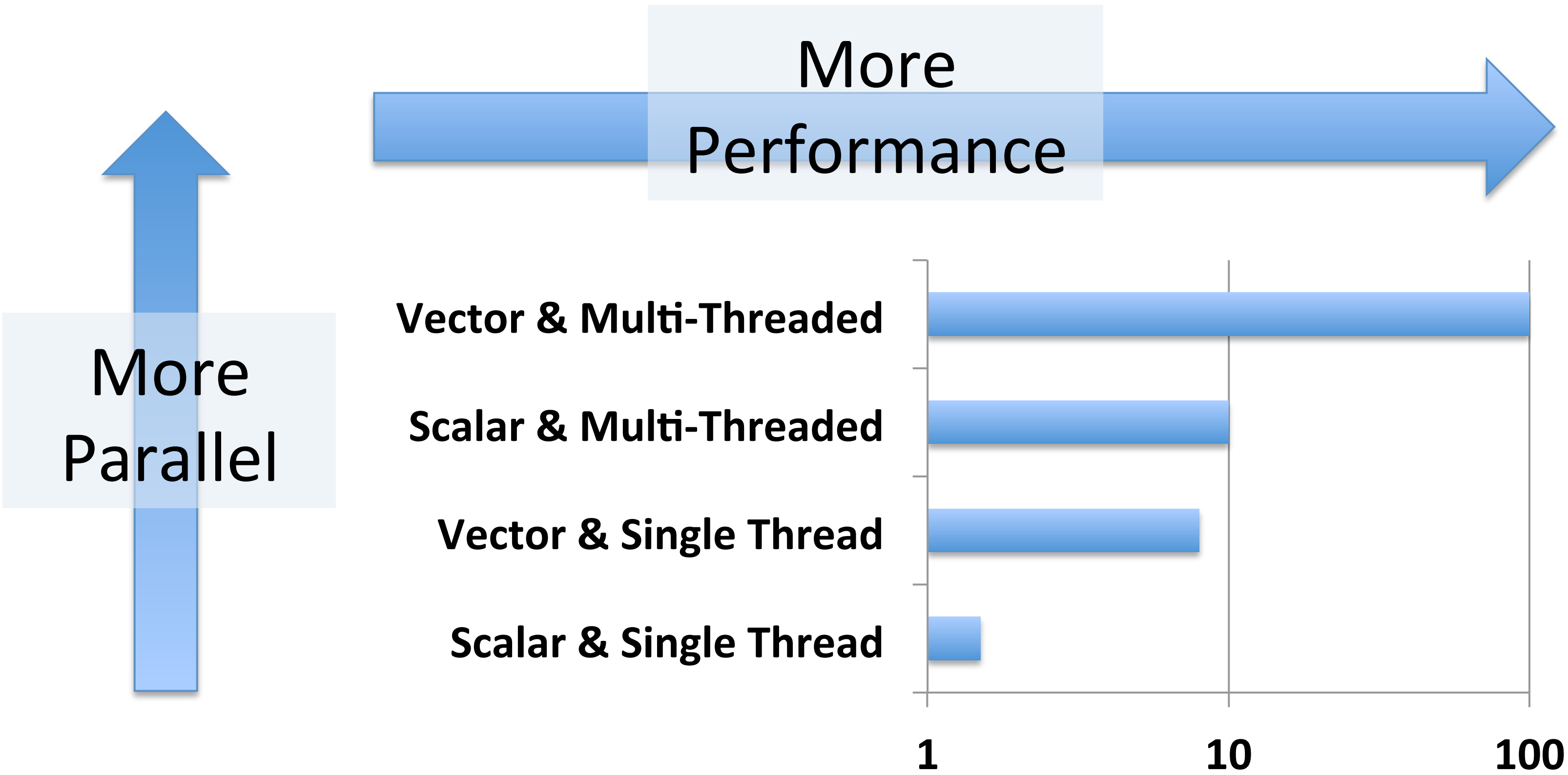
- When all CPU component work at maximum speed that is called *peak of performance*
  - Tech-spec normally describe the theoretical peak
  - Benchmarks measure the real peak
  - Applications show the real performance value
- CPU performance is measured as:
  - Floating point operations per seconds GFLOP/s
- But the real performance is in many cases mostly related to the memory bandwidth (GBytes/s)

Multi-core system Vs Serial Programming



Xeon E5650 hex-core processors (12GB - RAM)

Threading and Vectorization



Aid Automatic Vectorization

- Avoid Data Dependences
- Data Alignment (memalign)
- Aliasing
- Avoid Conditional Statements
- Analyze compiler report
  – -qopt-report=5 -qopt-report-phase=vec
- Add Compiler options:
  – -axMIC-AVX512 (KNL) -march=core-avx2 (BWD)

Conclusion

- Development of today computer architecture is based on increasing complexity: the software crisis!!!
- Scientific software developers for high-performance computing need to master this complexity:
  - avoid useless instructions, branches (if possible) and expensive OP
  - enhance data locality and data reuse (memory throughput)
  - aid the compiler for automatic vectorization
  - use compiler optimization (-O3)
  - make use of computer libraries for HPC