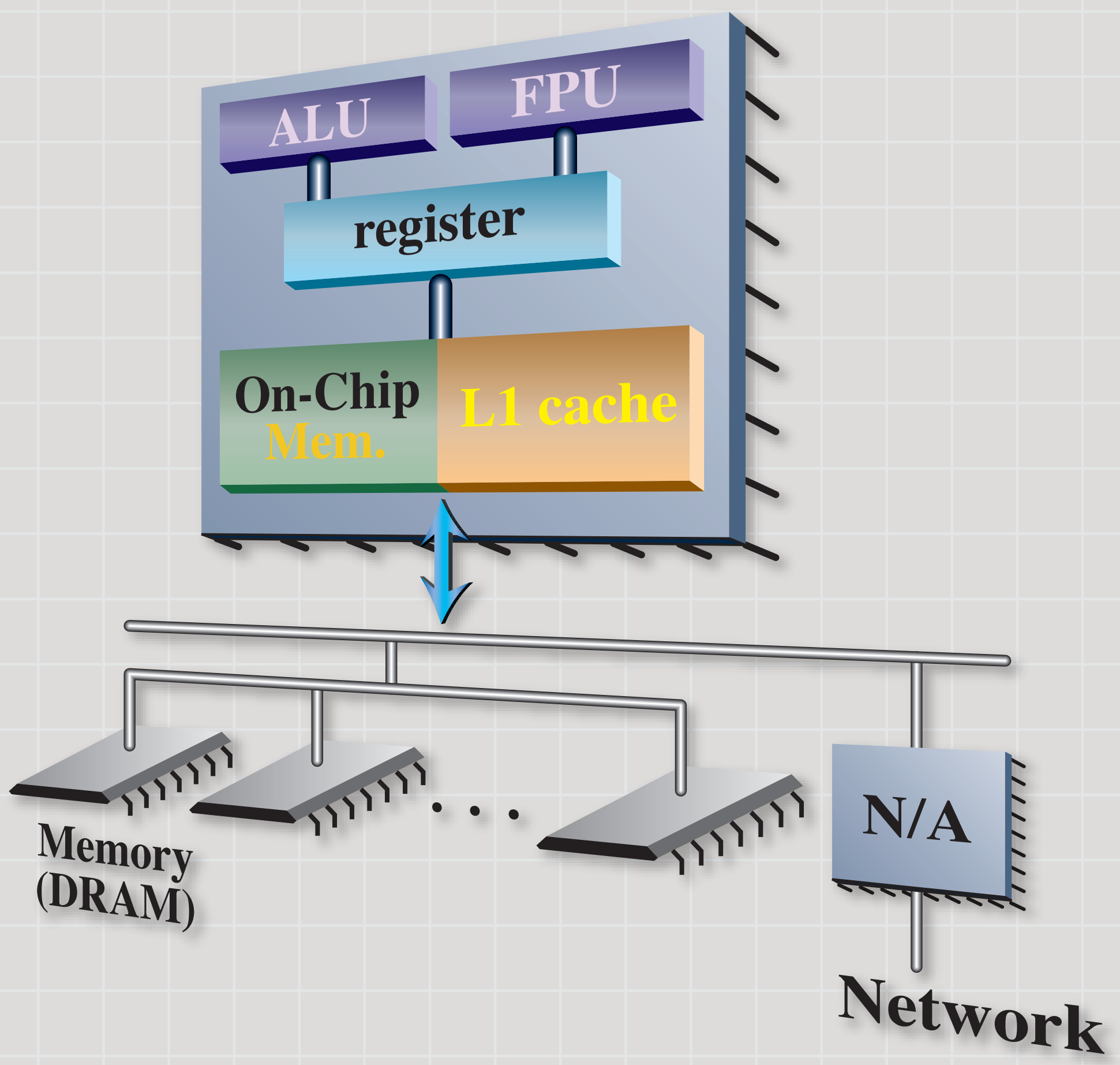
SCIMA:□Software Controlled Integrated □□□□□□Memory Architecture for HPC

Background

- Memory wall problem
- Conventional Cache is not good in HPC
  - □•unwilling line conflict
  - □•fixed size of Off-Chip Memory access

Solution: *SCIMA* *(Software Controlled Integrated Memory Architecture)*

- Strategy: software controllability
- Addressable **On-Chip Memory** in addition to conventional cache
  - □• On-Chip Memory and cache are reconfigurable
- Explicit data transfer between On-Chip Memory and Off-Chip Memory by **page-load/page-store** instruction
  - □• Burst transfer and stride transfer are supported
  - •• SMP configuration is supported



Overview of SCIMA

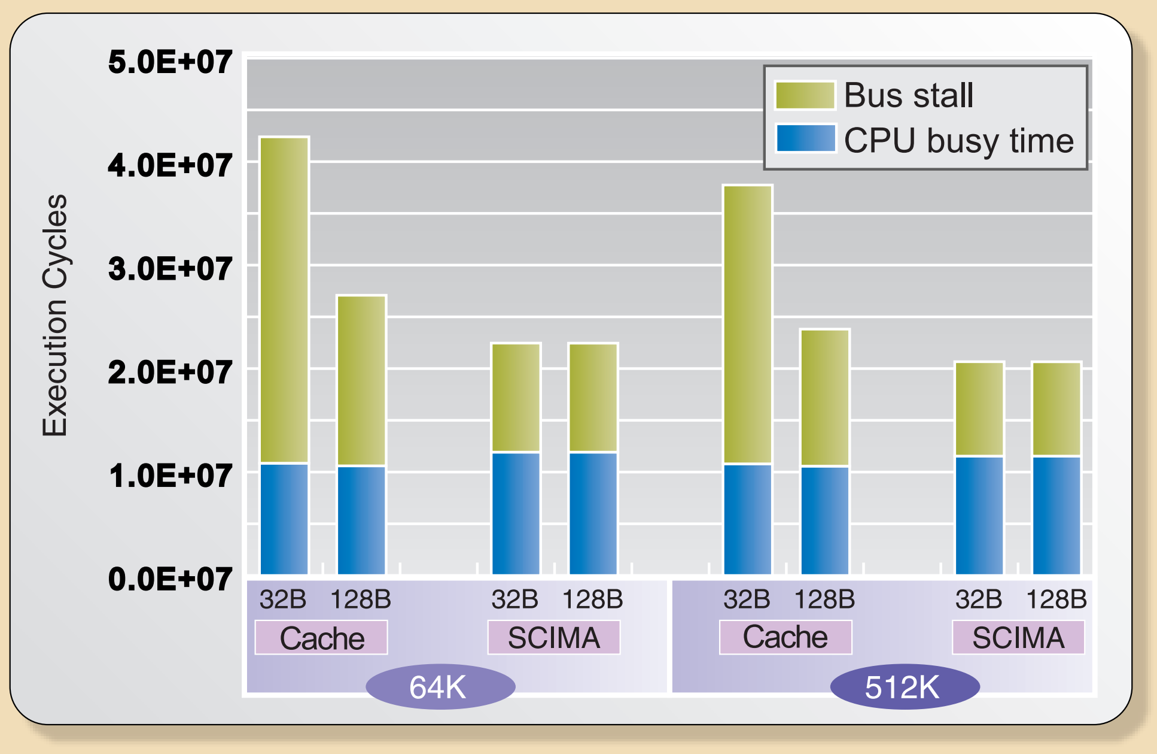
Advantages of SCIMA

Tb: CPU busy time *Tl:* Latency stall time *Tt:* Throughput stall time

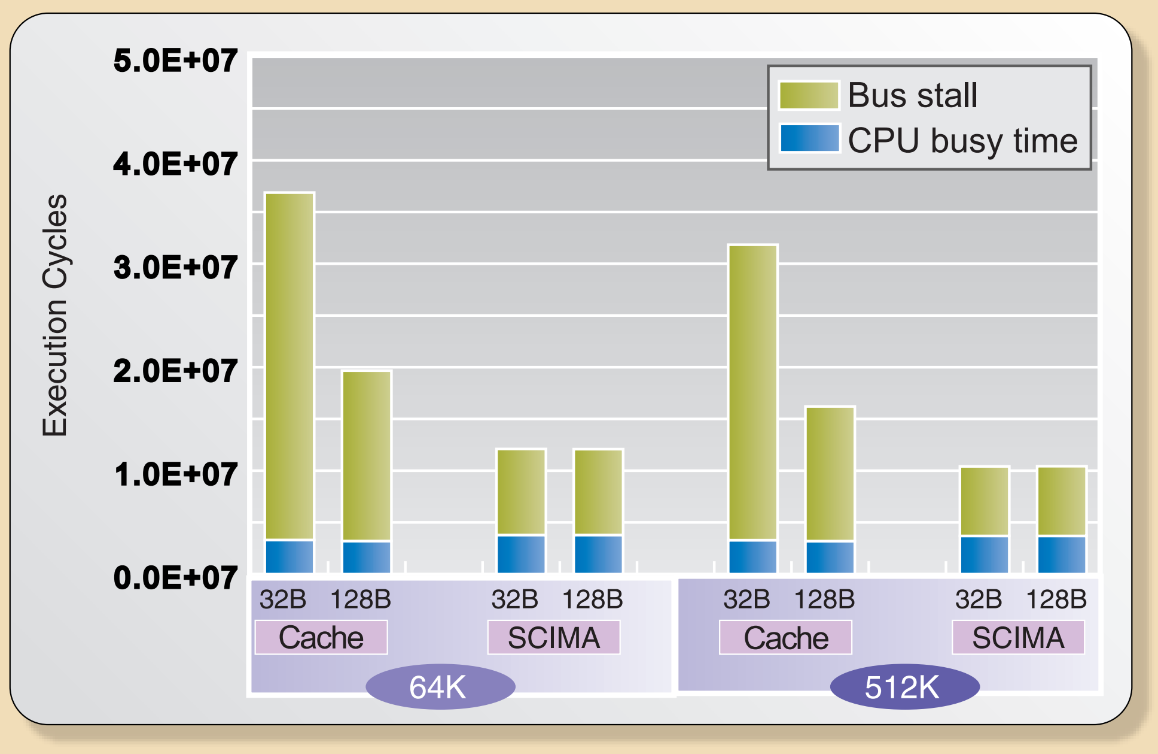
On-Chip Memory Features	*Tb*	*Tl*	*Tt*
□ software controllability	-	-	↓
□ page-load/page-store(burst)	↑	↓	-
□ page-load/page-store(stride)	↑	↓	↓
□ scheduling for page-load/page-store	-	↓	-
Latency Tolerating Techniques of Cache	***Tb***	***Tl***	***Tt***
□ larger cache line	-	↓	↑
□ lock-up free cache	-	↓	↑
□ cache prefetching	↑	↓	↑

Evaluation Results

QCD



1CPU



4CPUs

Throughput ratio = 2:1
Latency = 40 cycles

Throughput Ratio = ratio between On-Chip and Off-Chip memory throughput
Latency = memory access latency for Off-Chip memory (latency for first data)

The performance of SCIMA scales much better than Cache in SMP configuration because of reduced Bus Stall.