System architecture

Jakub Yaghob

Martin Kruliš

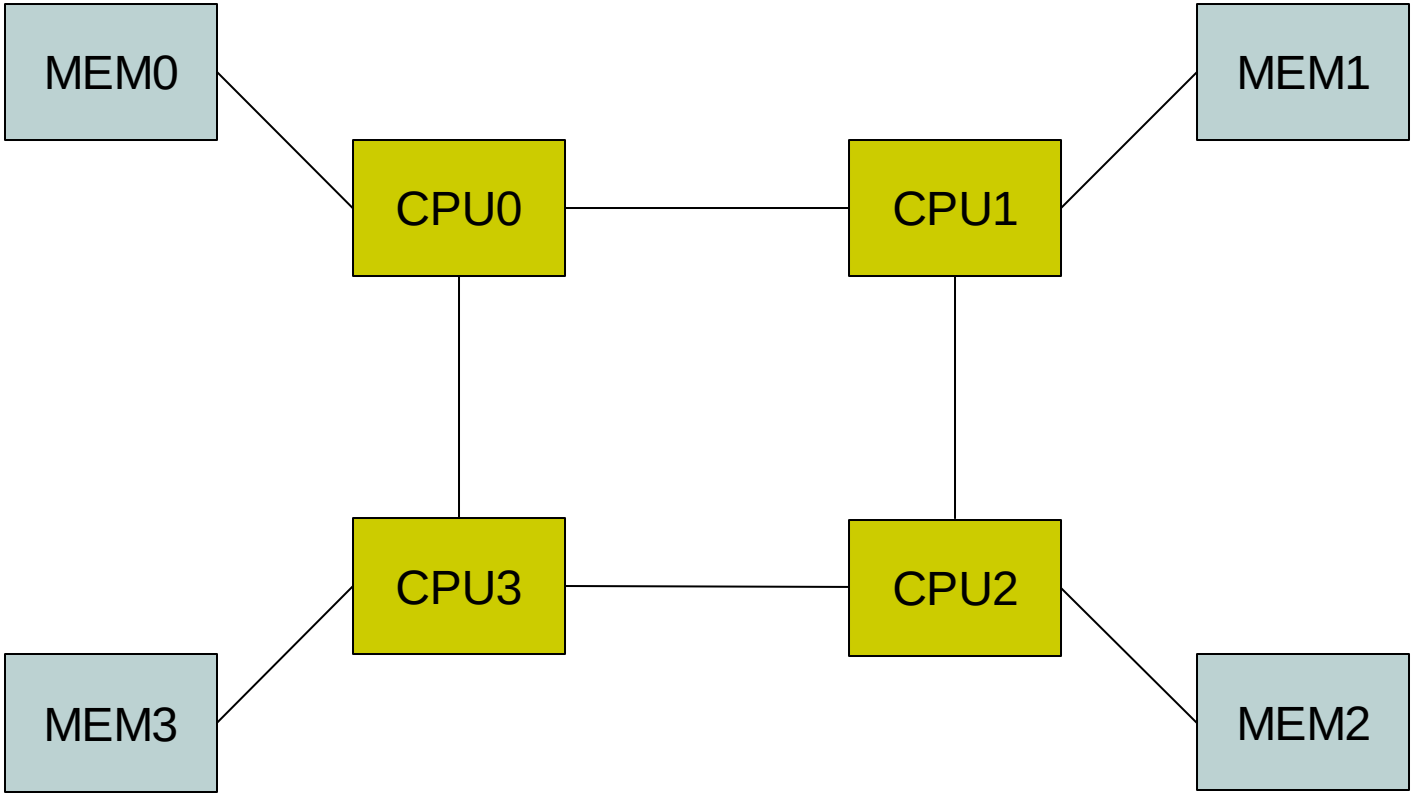
Terminology – 1

- Distributed systems
  - Host/node/system
    - One computer
  - Cluster
    - A set of nodes connected together by a network
    - Usually homogenous
  - Grid
    - A set of clusters connected by internet

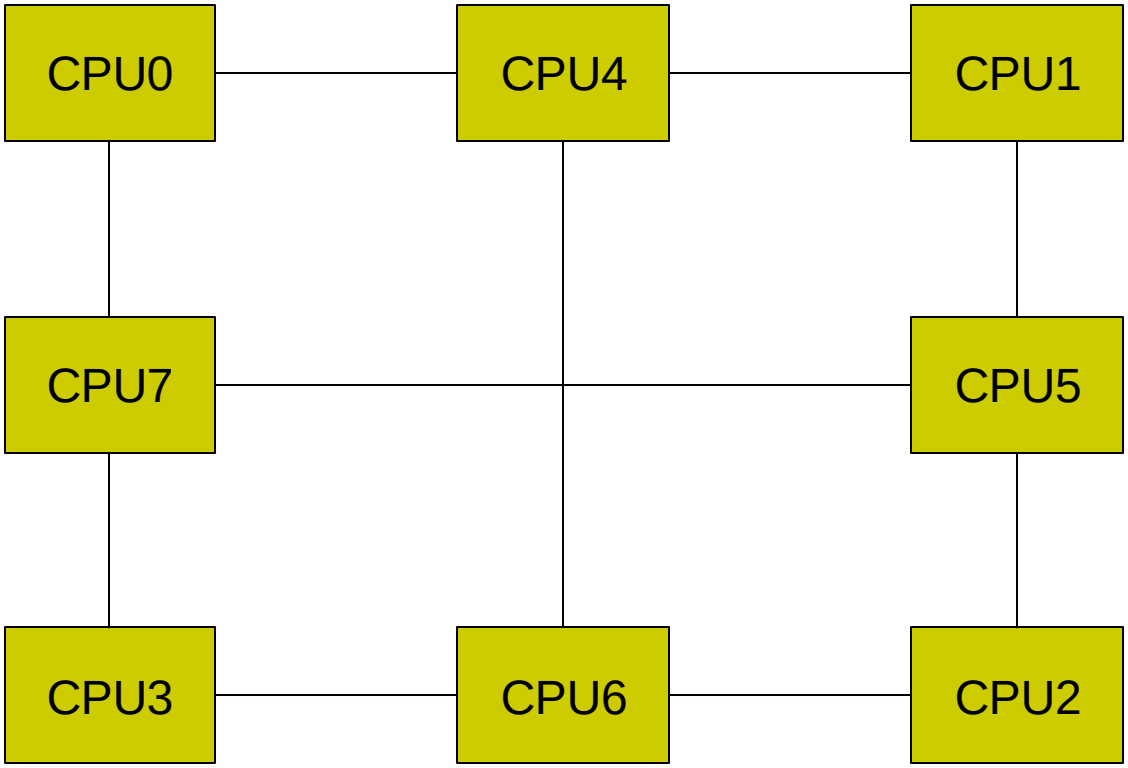
Terminology – 2

- One system
  - Socket/package
    - Physical CPU
    - Multiple sockets connected together by a high-speed connection
    - Cache hierarchy
    - Cores
  - NUMA node
    - Main memory
    - Can contain processing units
  - Core
    - Owns execution units
    - Shared registers
    - Contains logical CPUs
  - Logical CPU/thread
    - Processing unit
    - Executes instructions
    - Private registers
    - Hyper-threading

NUMA system – 4S

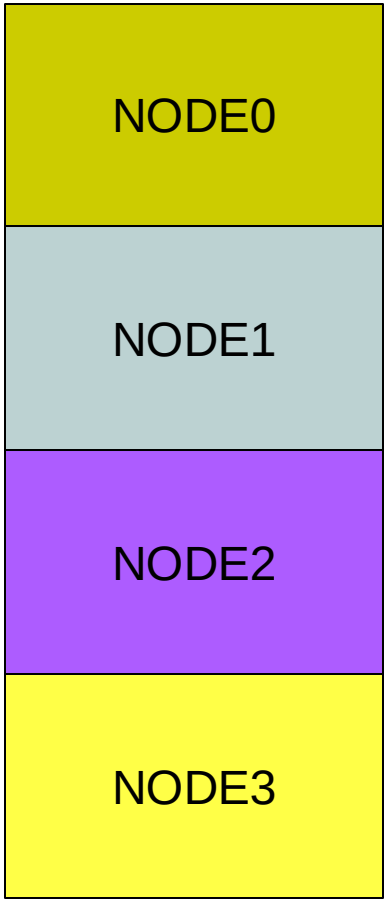


NUMA system – 8S

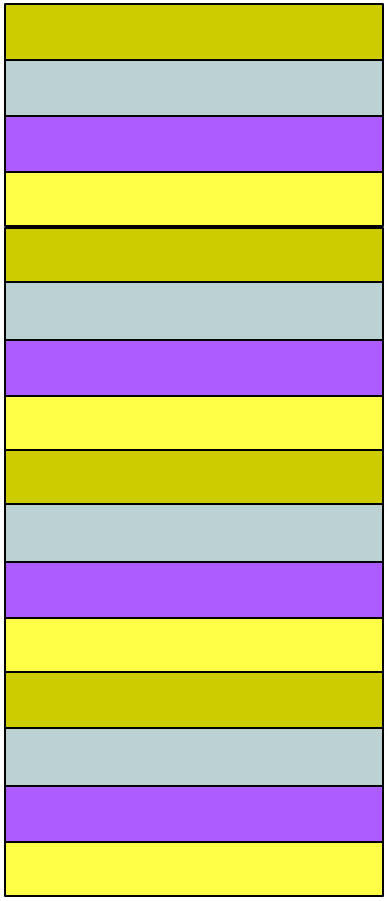


NUMA – physical memory layout

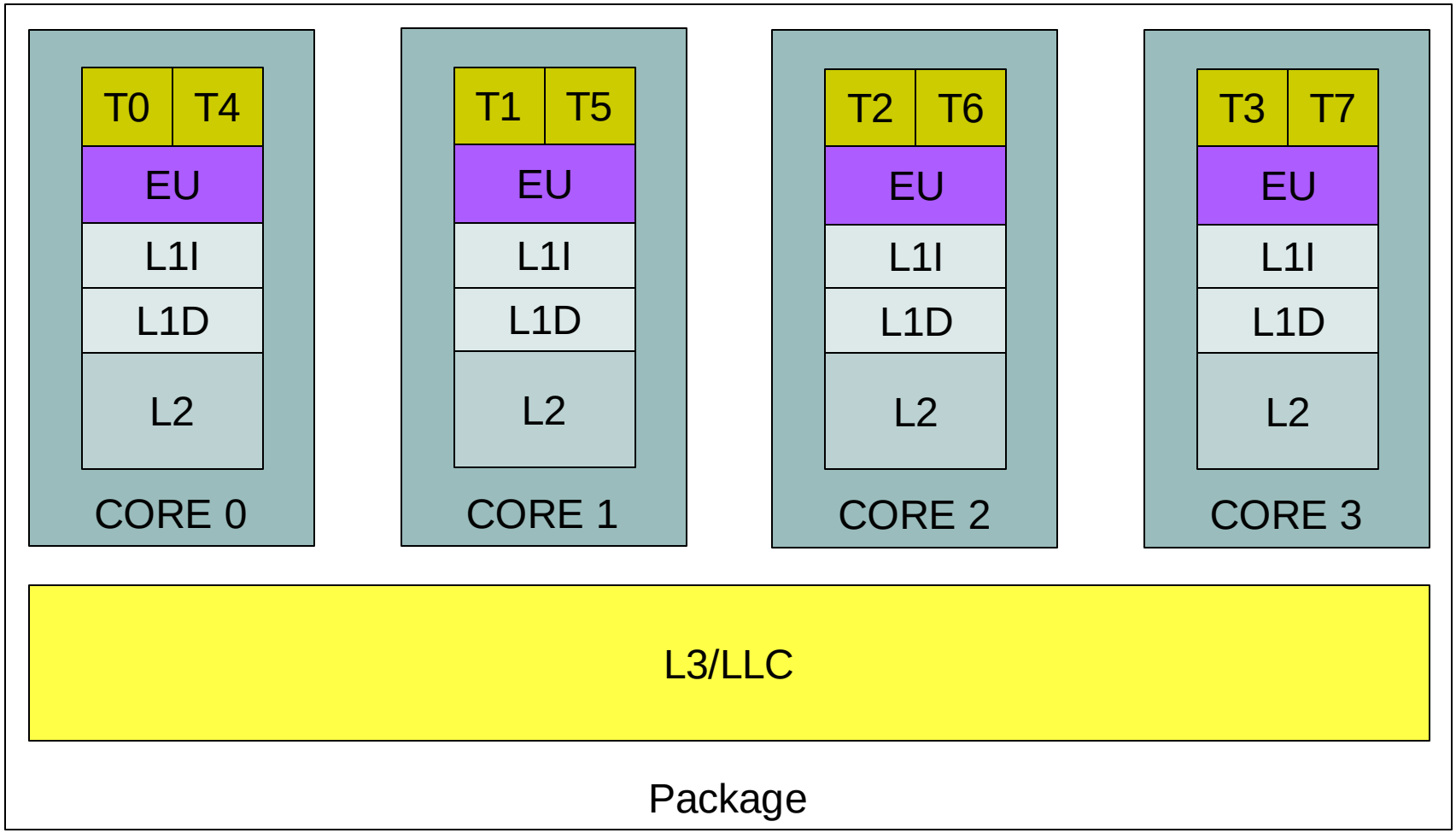
Block



Interleaving



Simplified package architecture



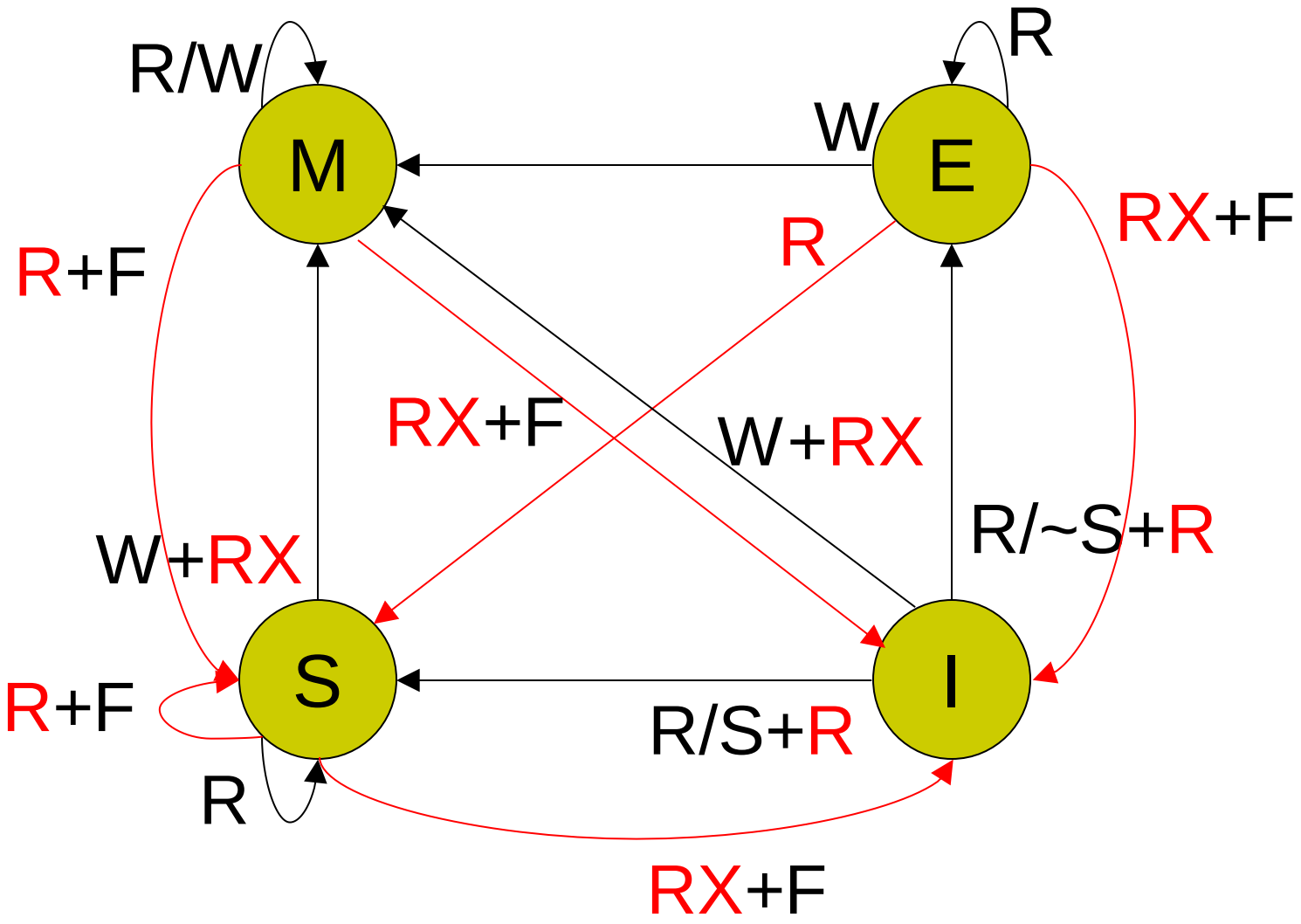
Cache terminology

- Cache line
  - Data transferred between memory and cache in atomic blocks
    - 64B
- Cache hit
  - Data load/store from/to a cache
- Cache line load
  - Cache line read from main memory
- Cache line flush
  - Cache line stored to main memory
- Cache miss
  - A cache line is selected for eviction
  - If it is modified, cache line will be flushed
  - The cache line is loaded
- False sharing
  - Private data of different threads in the same cache line

Cache coherency

- Coherency inside the package
  - Inclusive x exclusive caches
- Coherency between packages
  - ccNUMA
  - MESI protocol
    - Modified, Exclusive, Shared, and Invalid
    - Snooping
- Cache line ping-pong
  - Moving cache line among caches/packages in rapid succession

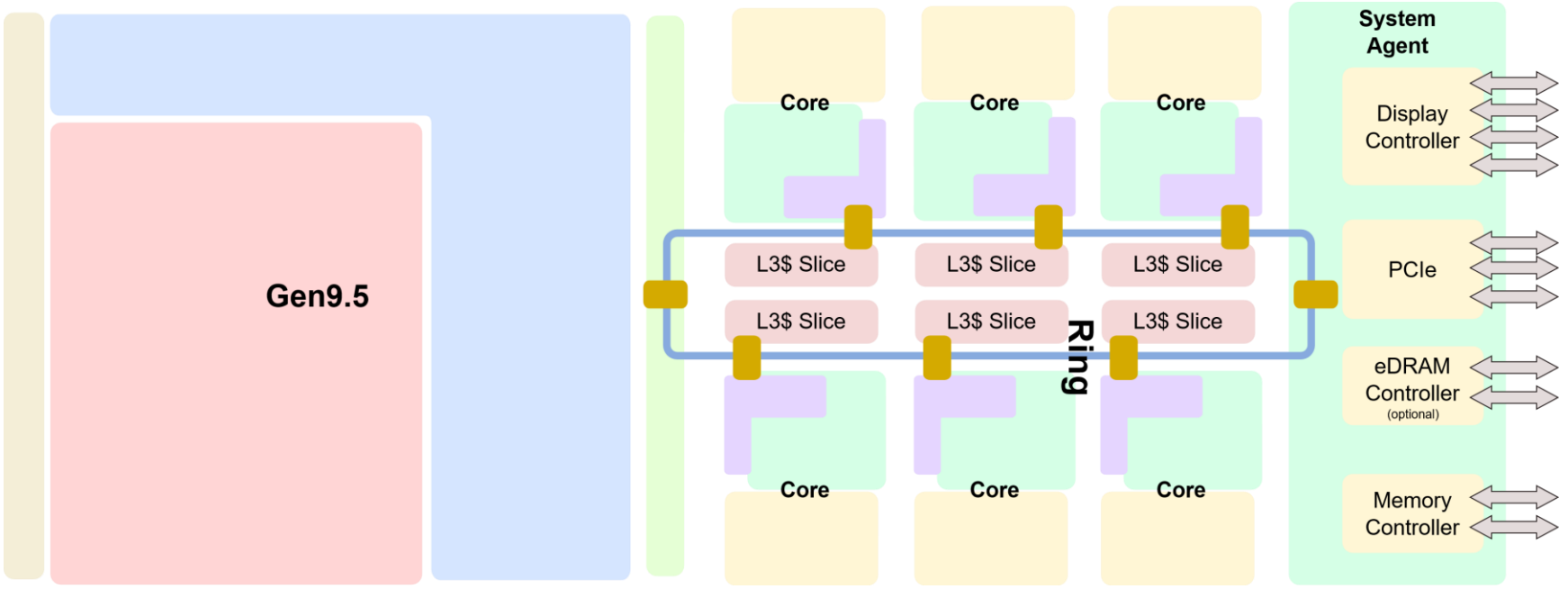
MESI protocol



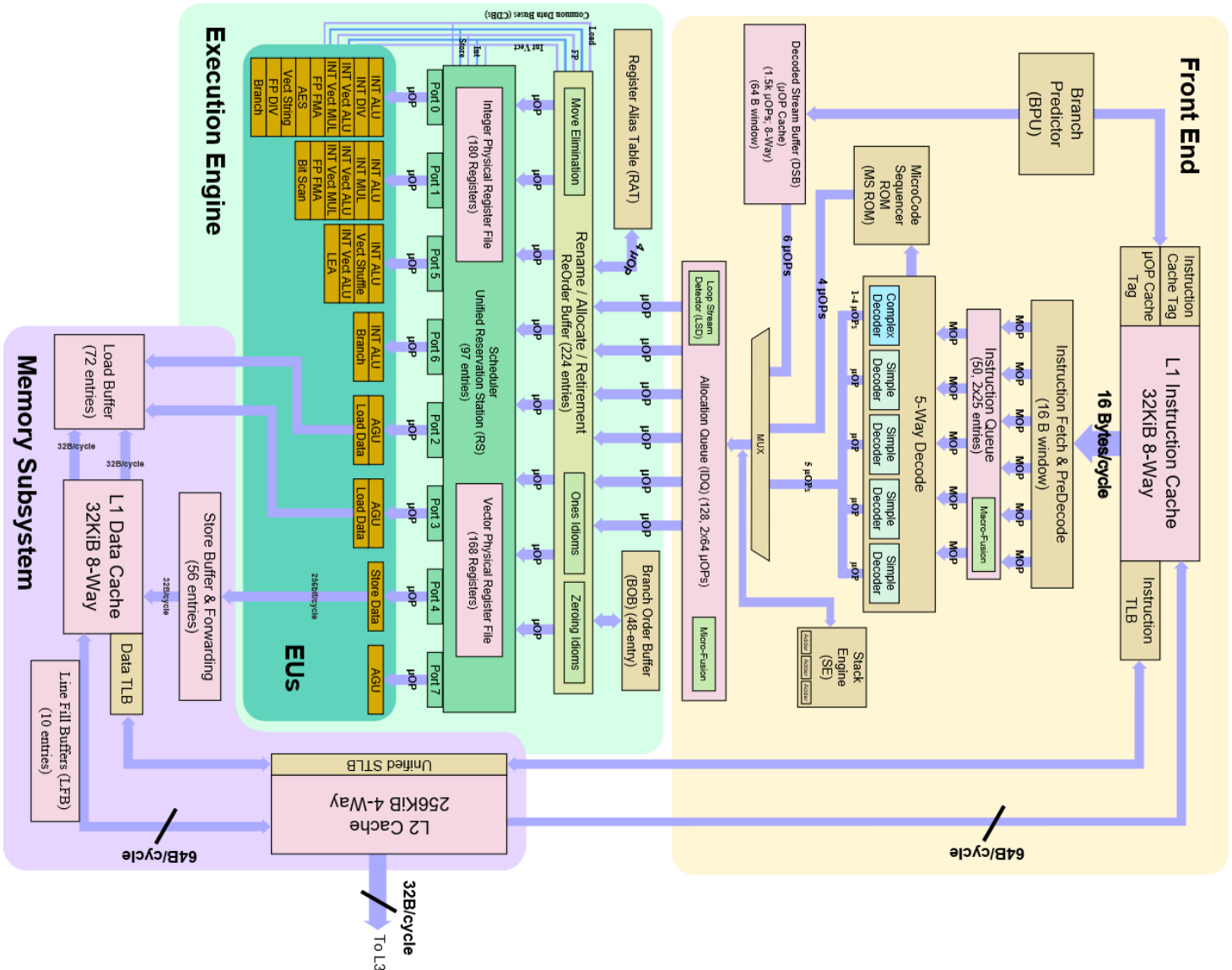
Latencies

Action	Cycles	Time (3GHz)
Local L1 cache hit	4	1-2 ns
Local L2 cache hit	14	4-6 ns
Branch misprediction	16	5-6 ns
Local L3 cache hit	40-75	12-40 ns
Mutex lock/unlock		75 ns
Remote L3	100-300	30-100 ns
Local memory		100 ns
Remote memory		100-300 ns
Send 1KB over FDR InfiniBand		900 ns
Send 1KB over 1GB Ethernet		20000 ns

Package schema



Core schema


Front End
Branch Predictor (BPU)
Instruction Cache Tag
µOP Cache Tag
L1 Instruction Cache 32KiB 8-Way
Instruction TLB
16 Bytes/cycle
Instruction Fetch & PreDecode (16 B window)
Instruction Queue (50, 2x25 entries)
Macro-Fusion
MOP
Decoded Stream Buffer (DSB) (µOP Cache) (1.5k µOPs, 8-Way) (64 B window)
6 µOPs
MicroCode Sequencer ROM (MS ROM)
4 µOPs
5-Way Decode
Complex Decoder
Simple Decoder
1-4 µOPs
5 µOPs
MUX
Stack Engine (SE)
Loop Stream Detector (LSD)
Allocation Queue (IDQ) (128, 2x64 µOPs)
Micro-Fusion
64B/cycle
Register Alias Table (RAT)
4 µOPs
Move Elimination
Rename / Allocate / Retirement ReOrder Buffer (224 entries)
Ones Idioms
Zeroing Idioms
Branch Order Buffer (BOB) (48-entry)
Execution Engine
Common Data Buses (CDBs)
Load
FP
Int Vect
Int
Store
Integer Physical Register File (180 Registers)
Scheduler Unified Reservation Station (RS) (97 entries)
Vector Physical Register File (168 Registers)
Port 0
Port 1
Port 5
Port 6
Port 2
Port 3
Port 4
Port 7
µOP
INT ALU
INT DIV
INT Vect ALU
INT Vect MUL
FP FMA
AES
Vect String
FP DIV
Branch
INT MUL
Bit Scan
Vect Shuffle
INT Vect LEA
Load Data
Store Data
AGU
EUs
256bit/cycle
Memory Subsystem
Load Buffer (72 entries)
32B/cycle
L1 Data Cache 32KiB 8-Way
Data TLB
Store Buffer & Forwarding (56 entries)
32B/cycle
Line Fill Buffers (LFB) (10 entries)
Unified STLB
L2 Cache 256KiB 4-Way
64B/cycle
32B/cycle
To L3