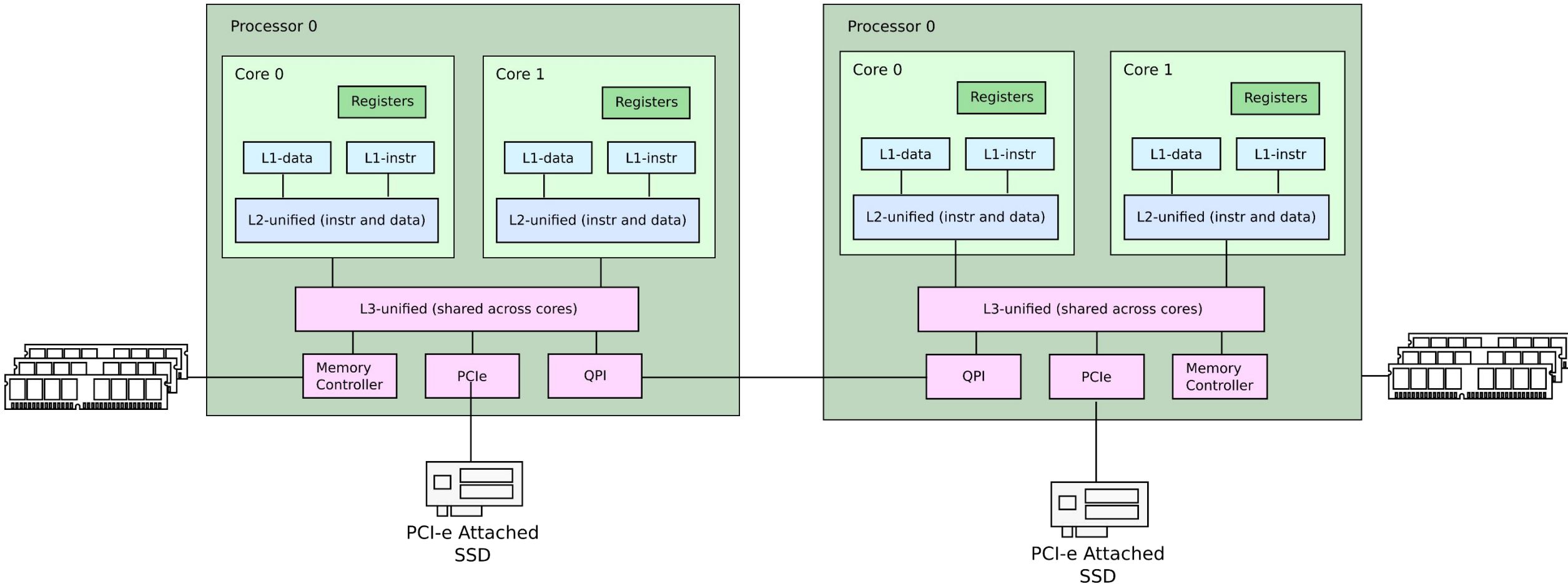
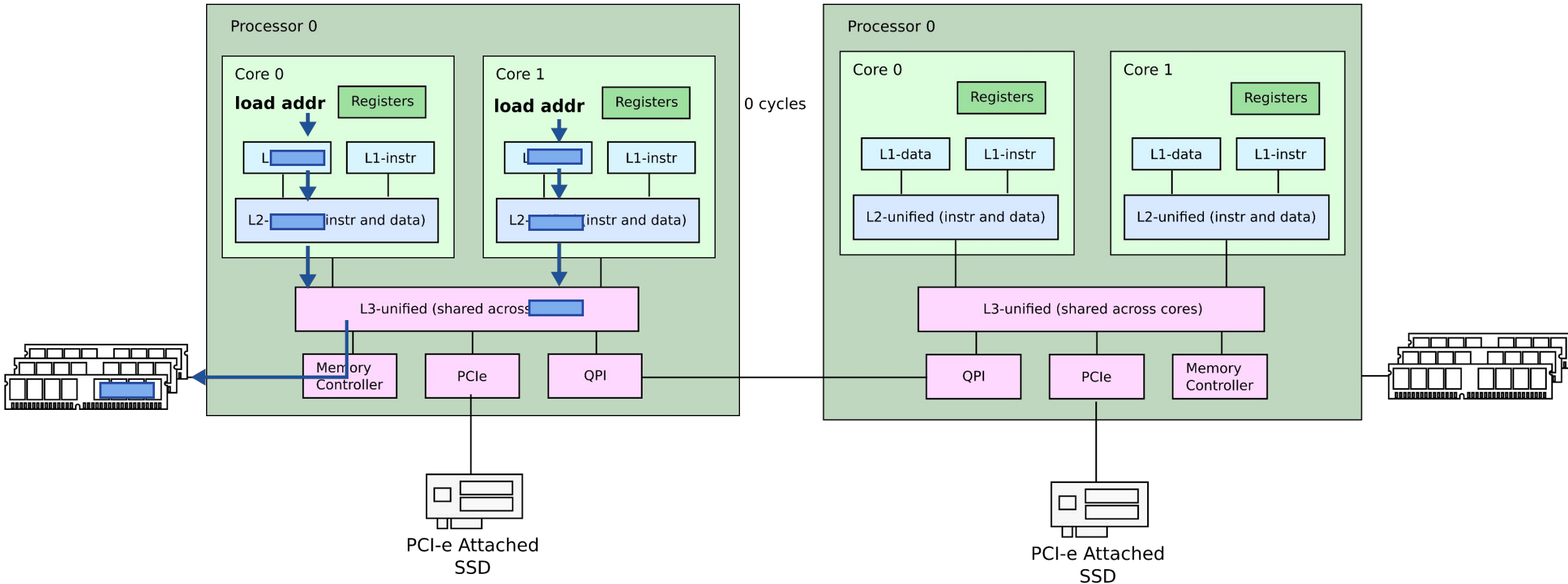


Intel Memory Hierarchy

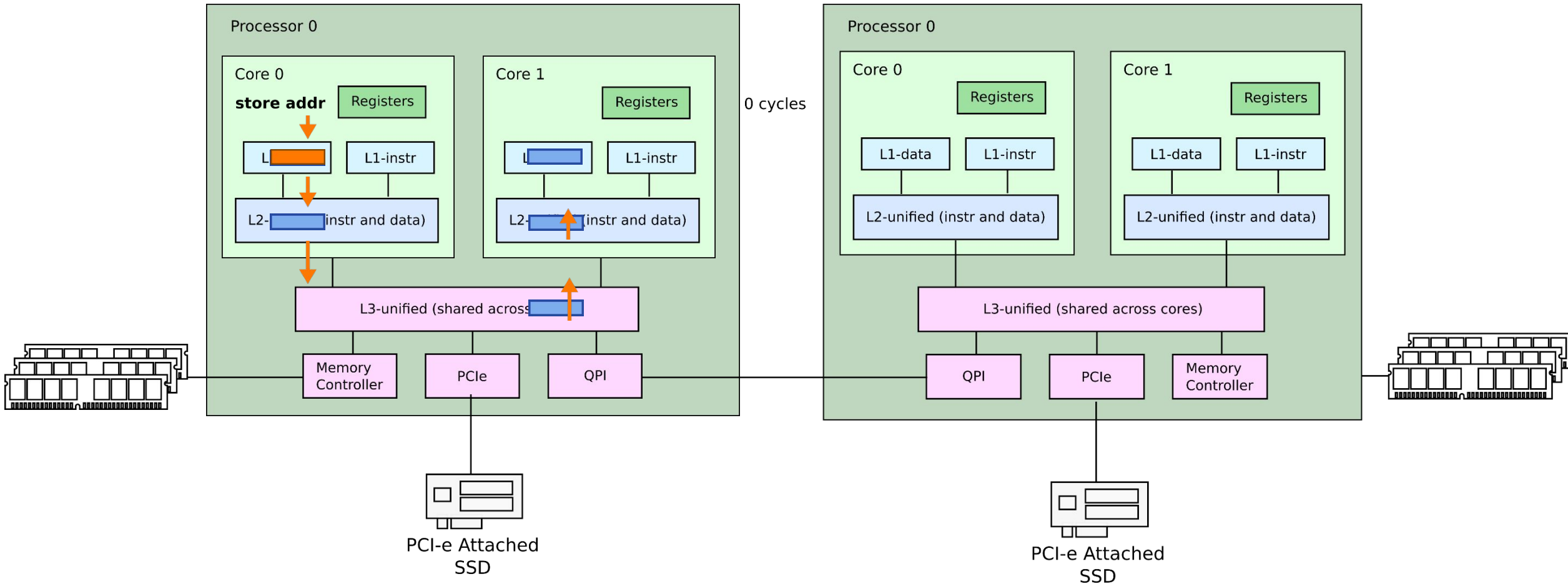
Processors, cores, memory and PCIe



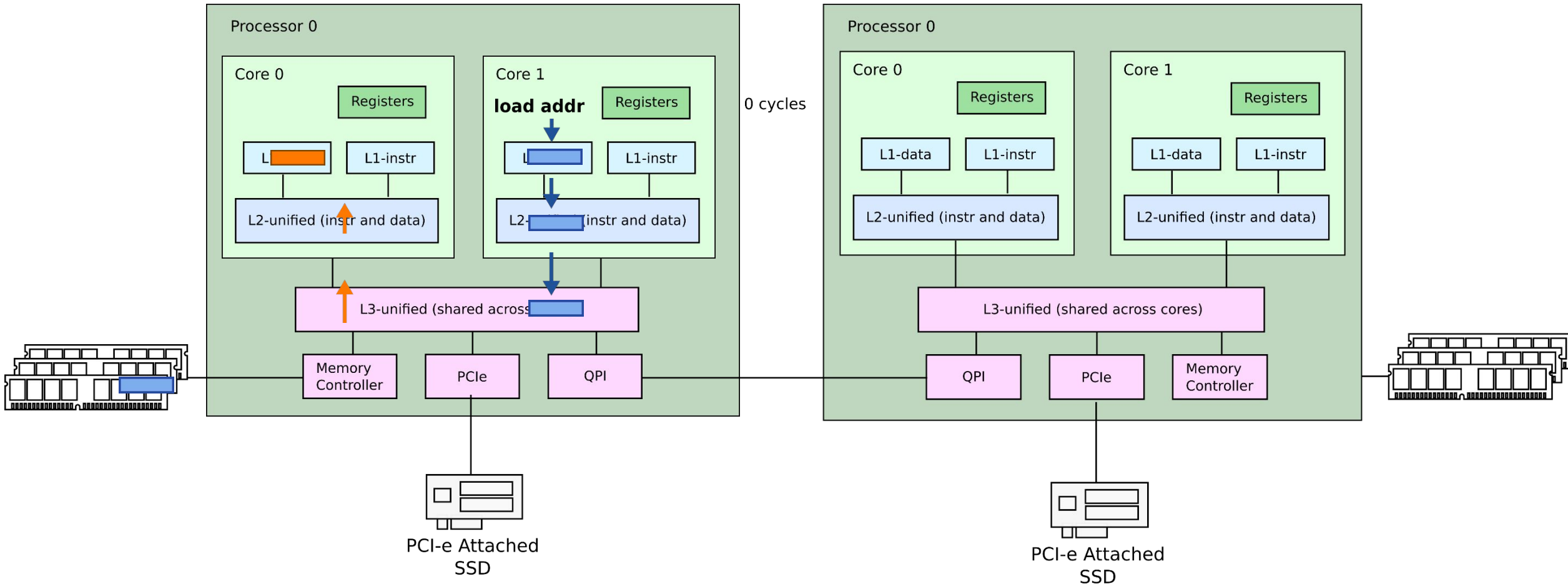
Caches (load)



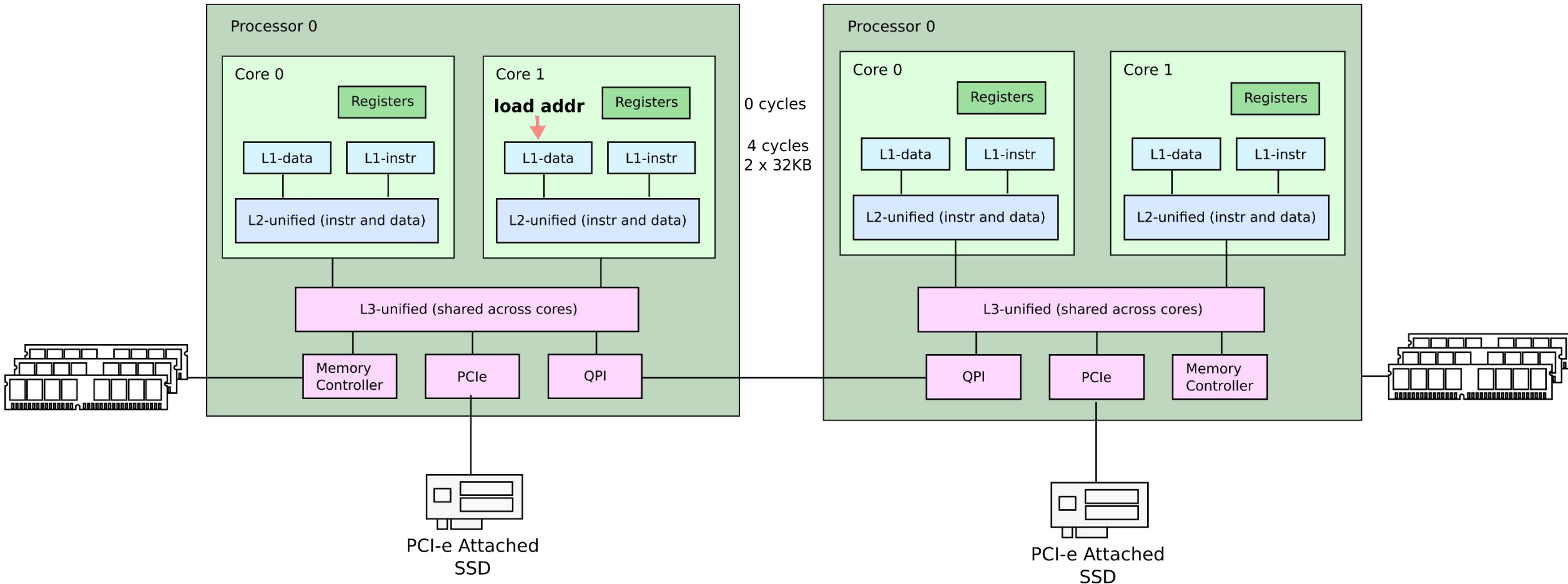
Cache-coherence (store)



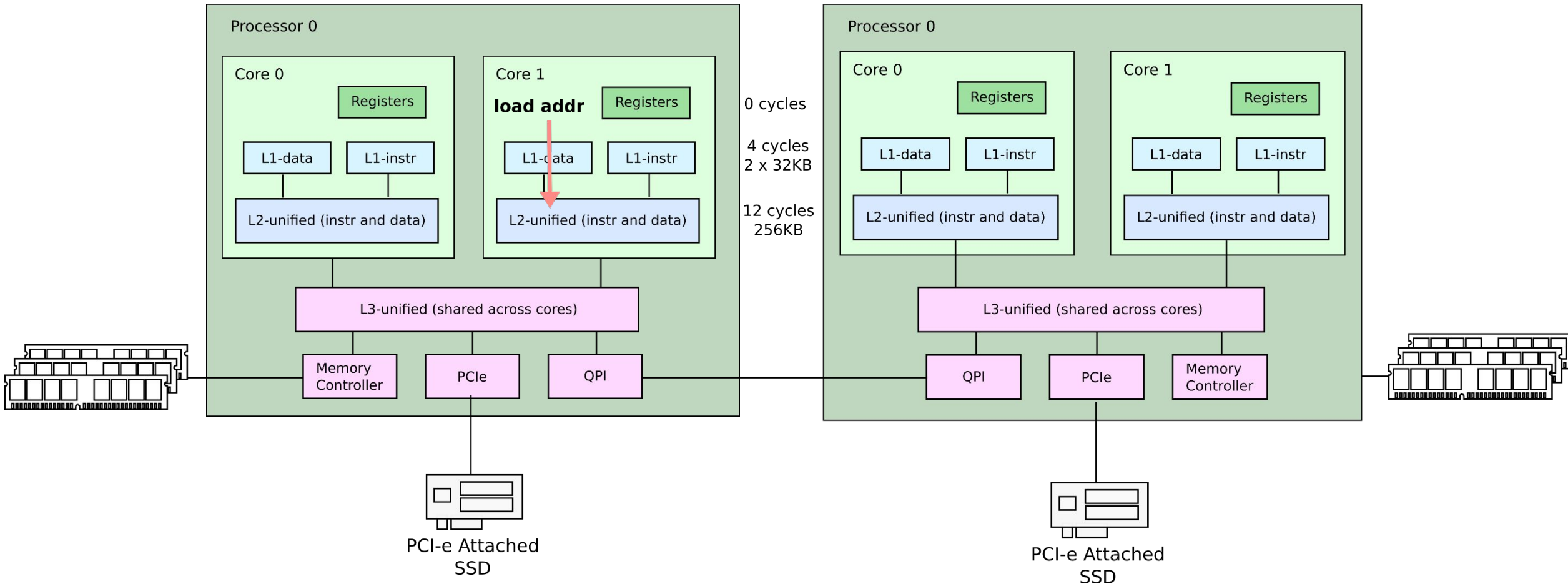
Cache-coherence (load of modified)



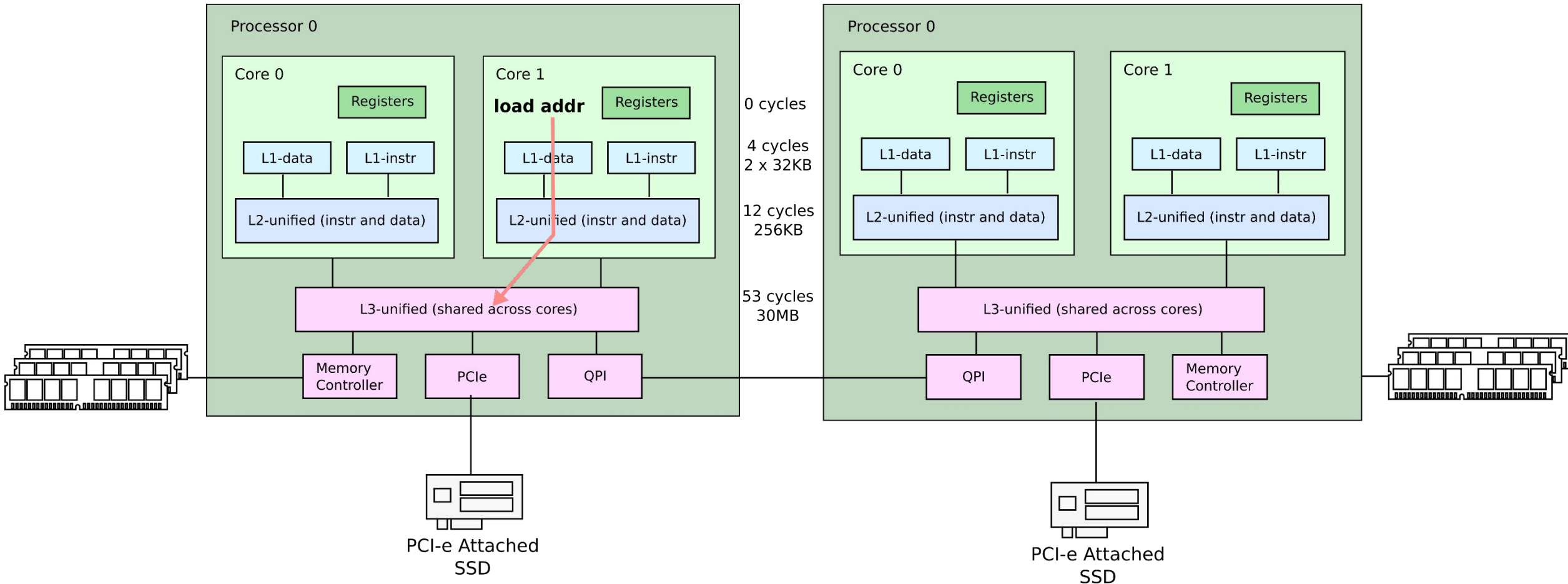
Latencies: load from local L1



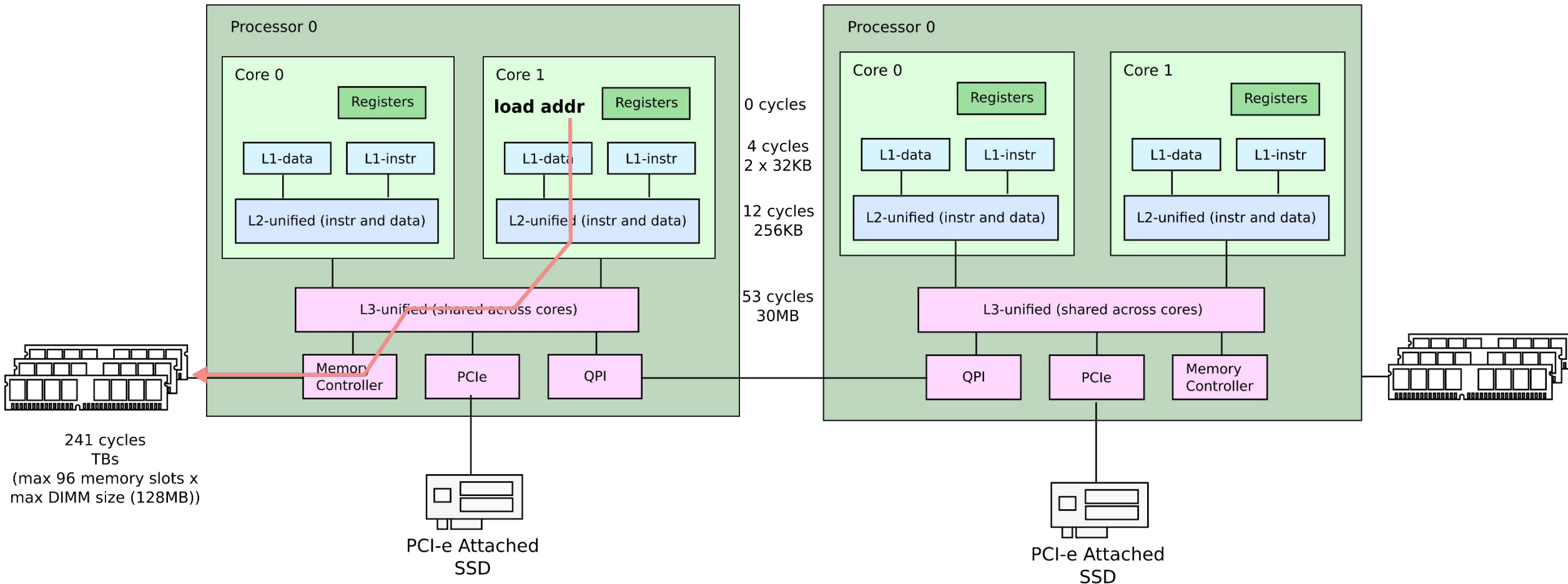
Latencies: load from local L2



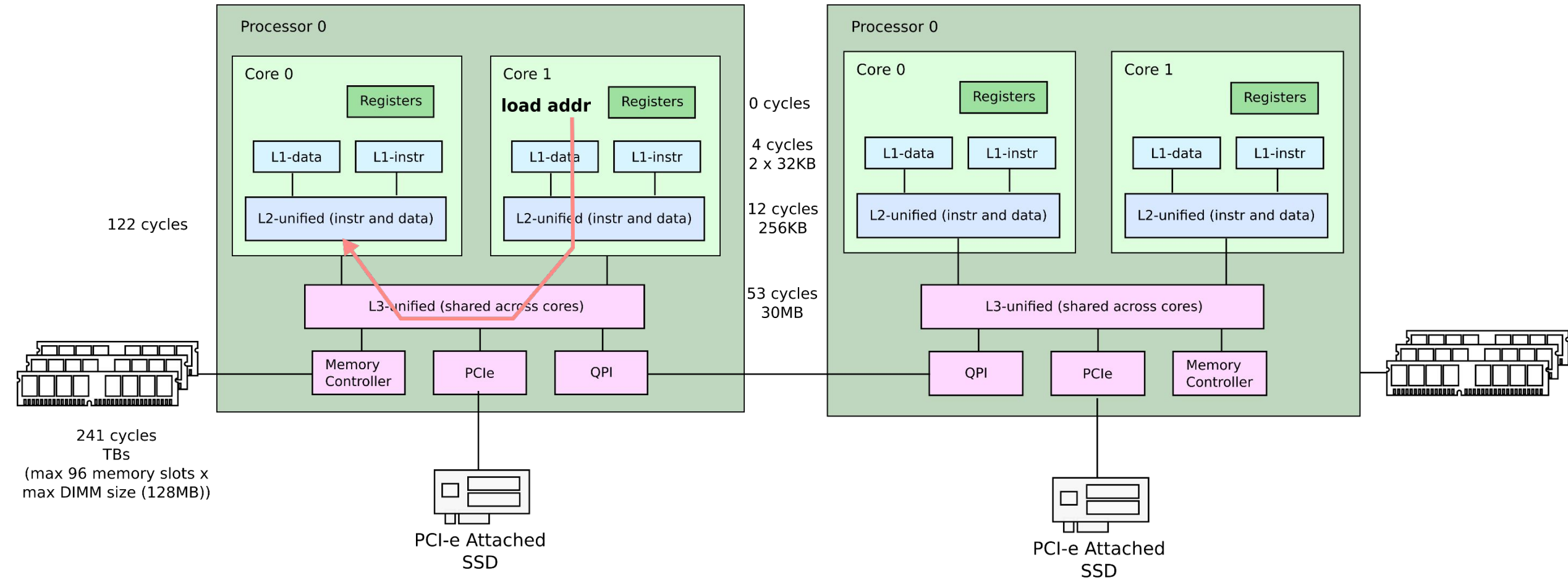
Latencies: load from local L3



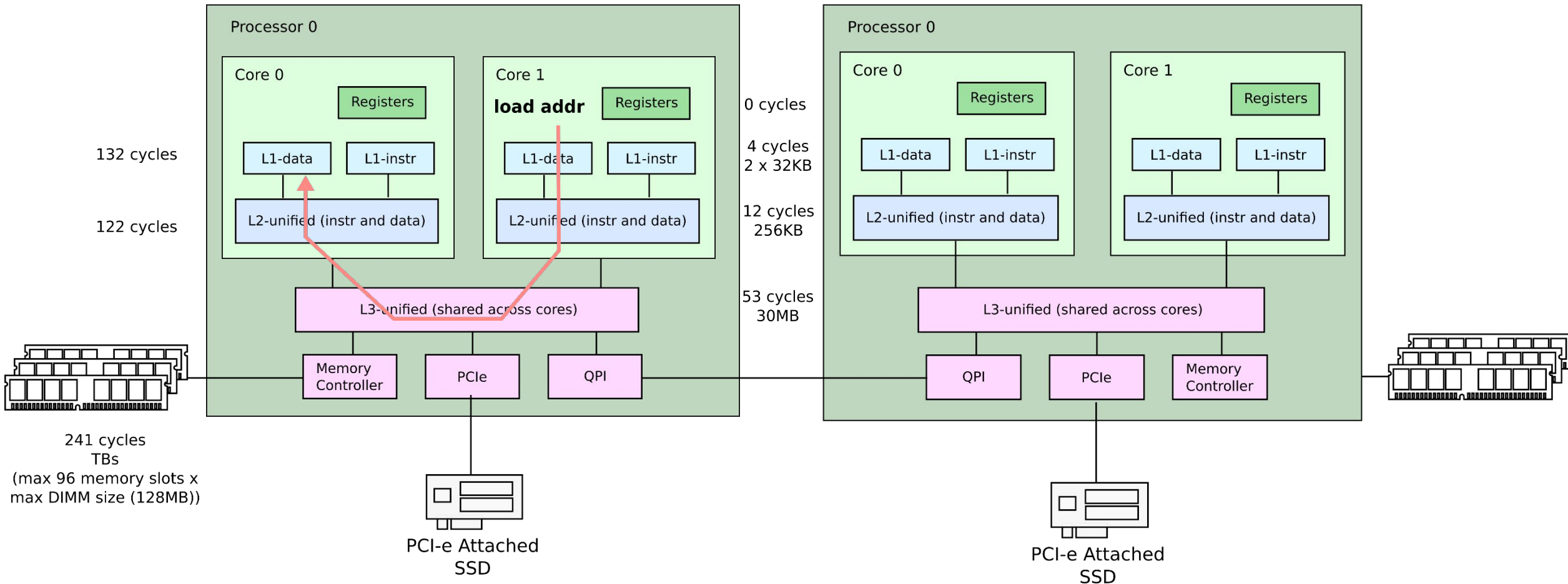
Latencies: load from local memory



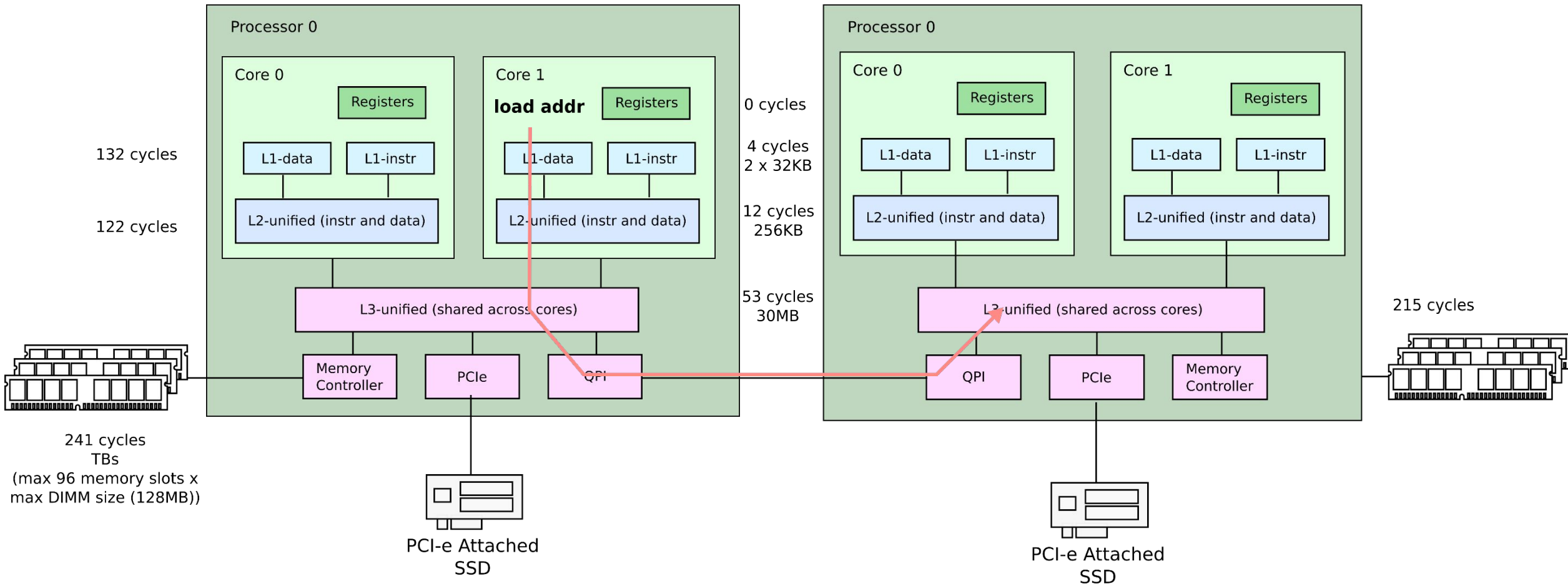
Latencies: load from same die core's L2



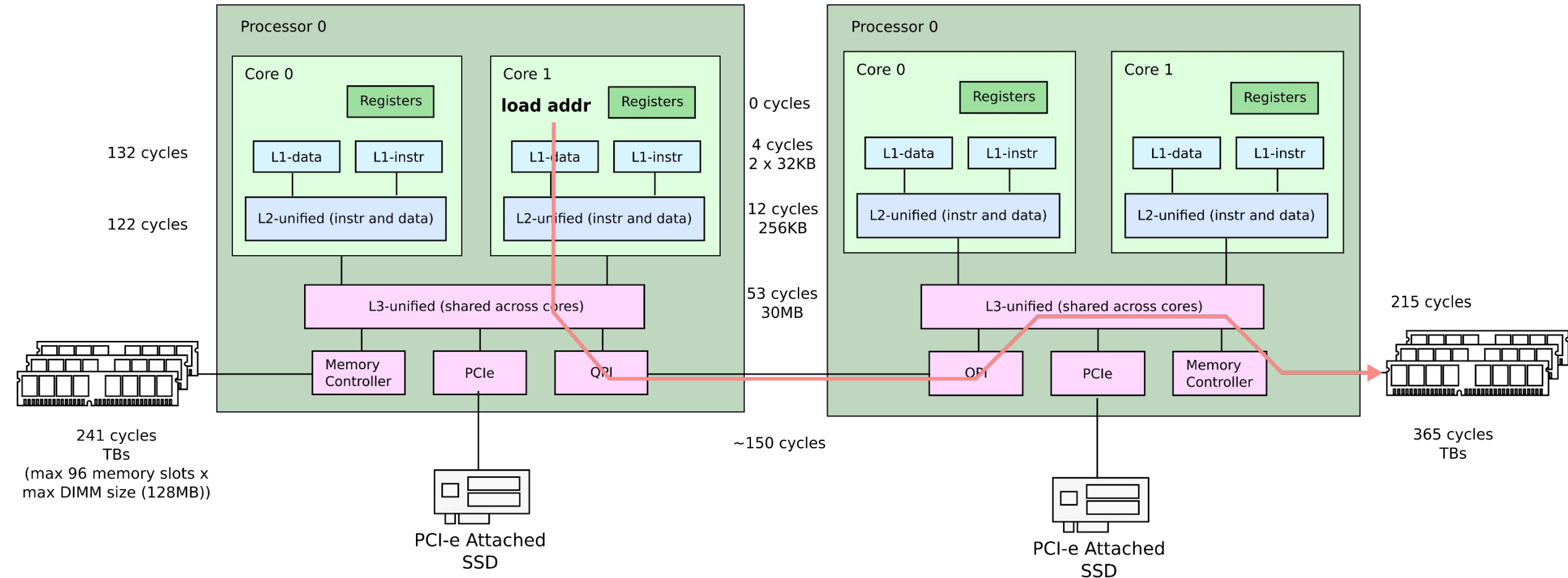
Latencies: load from same die core's L1



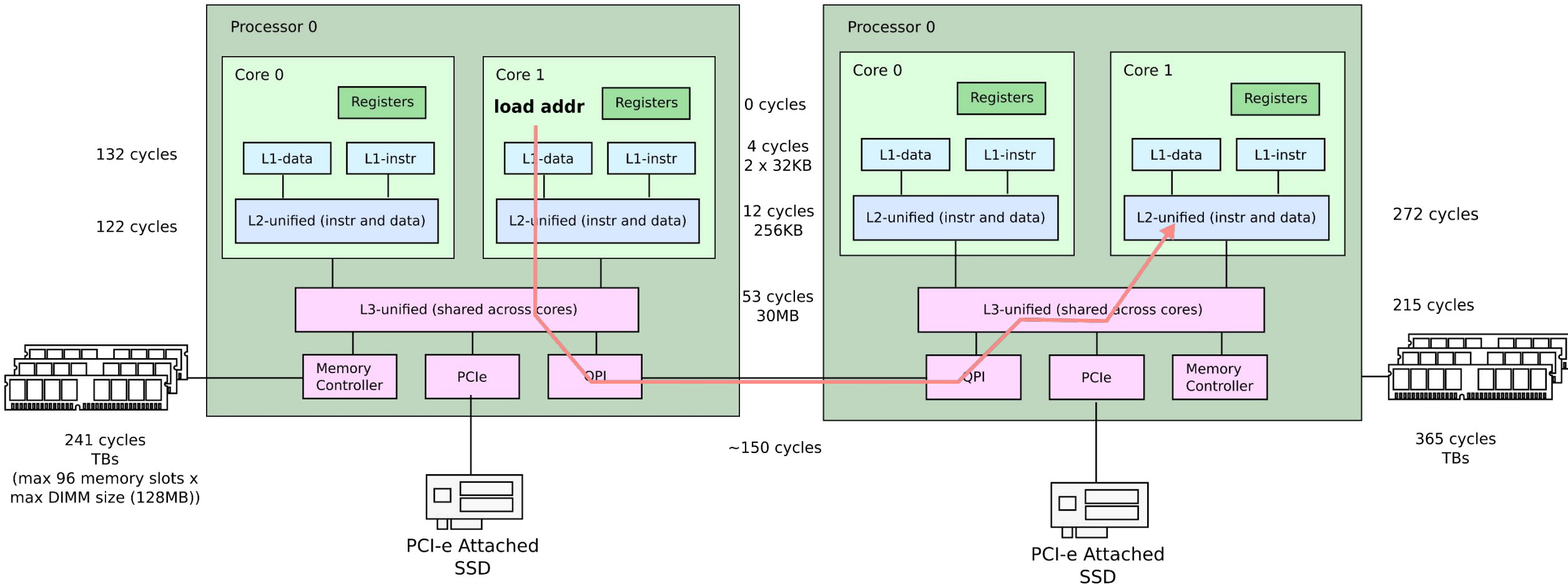
Latencies: load from remote L3



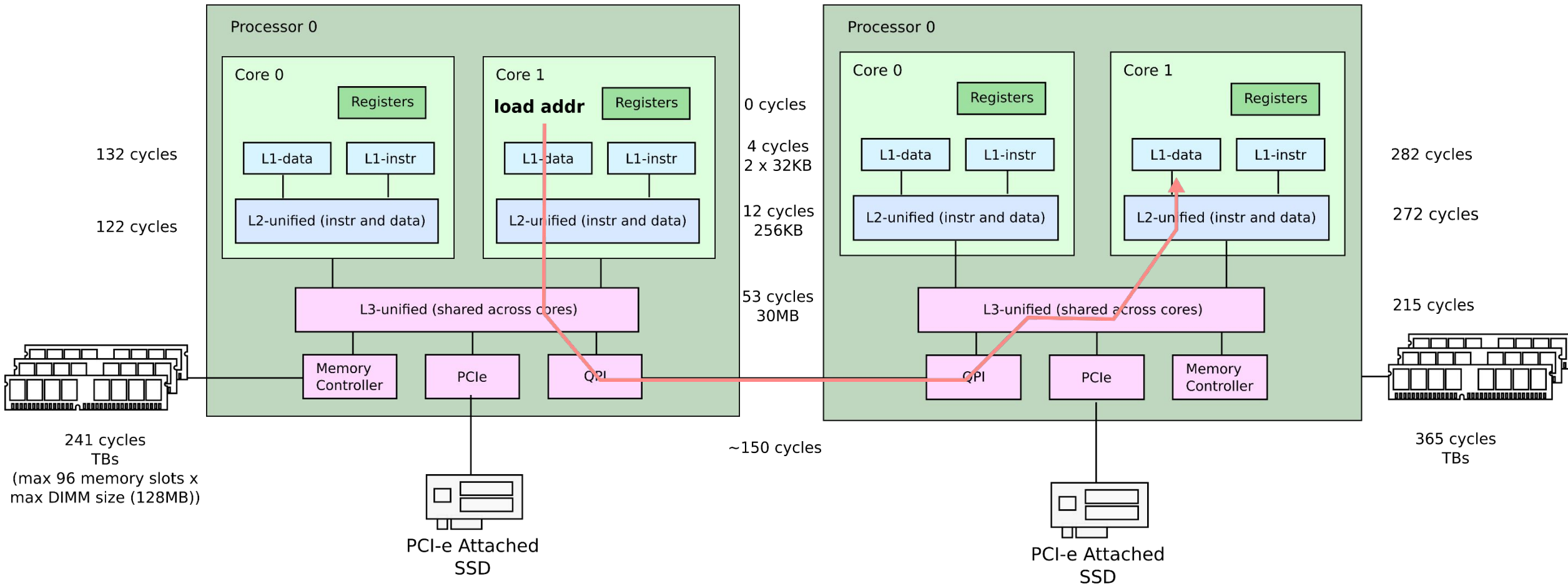
Latencies: load from remote memory



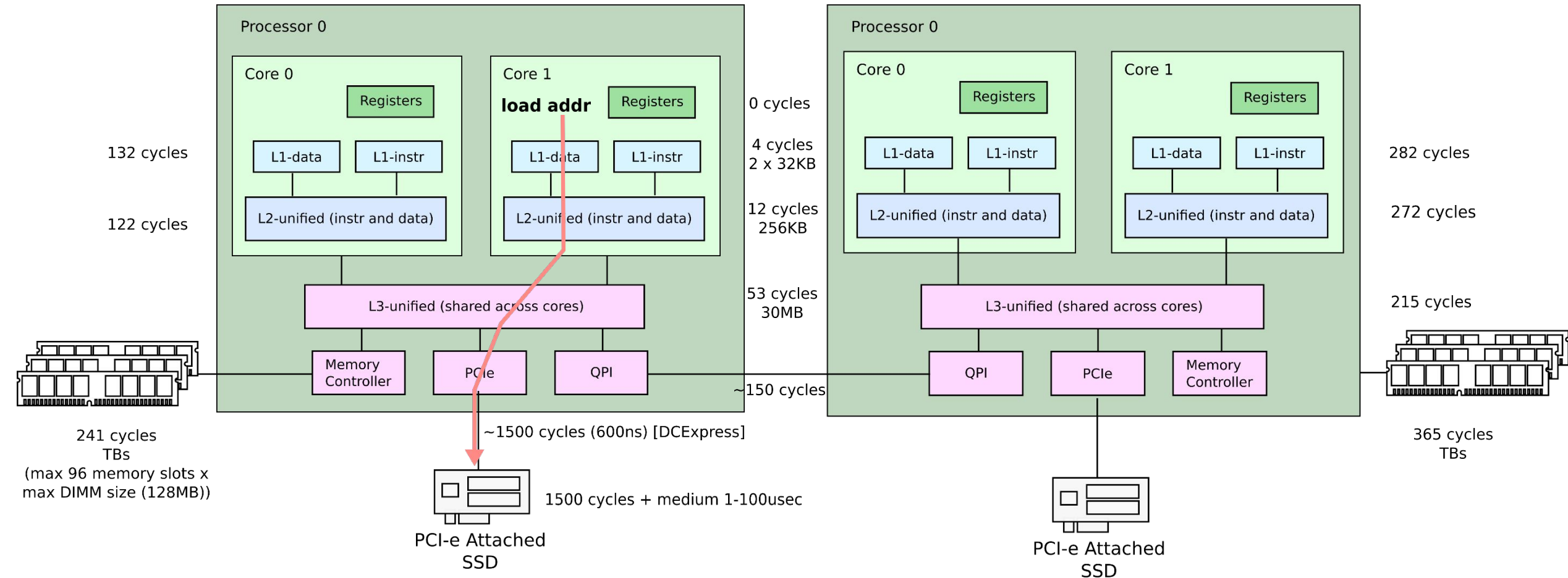
Latencies: load from remote L2



Latencies: load from remote L2



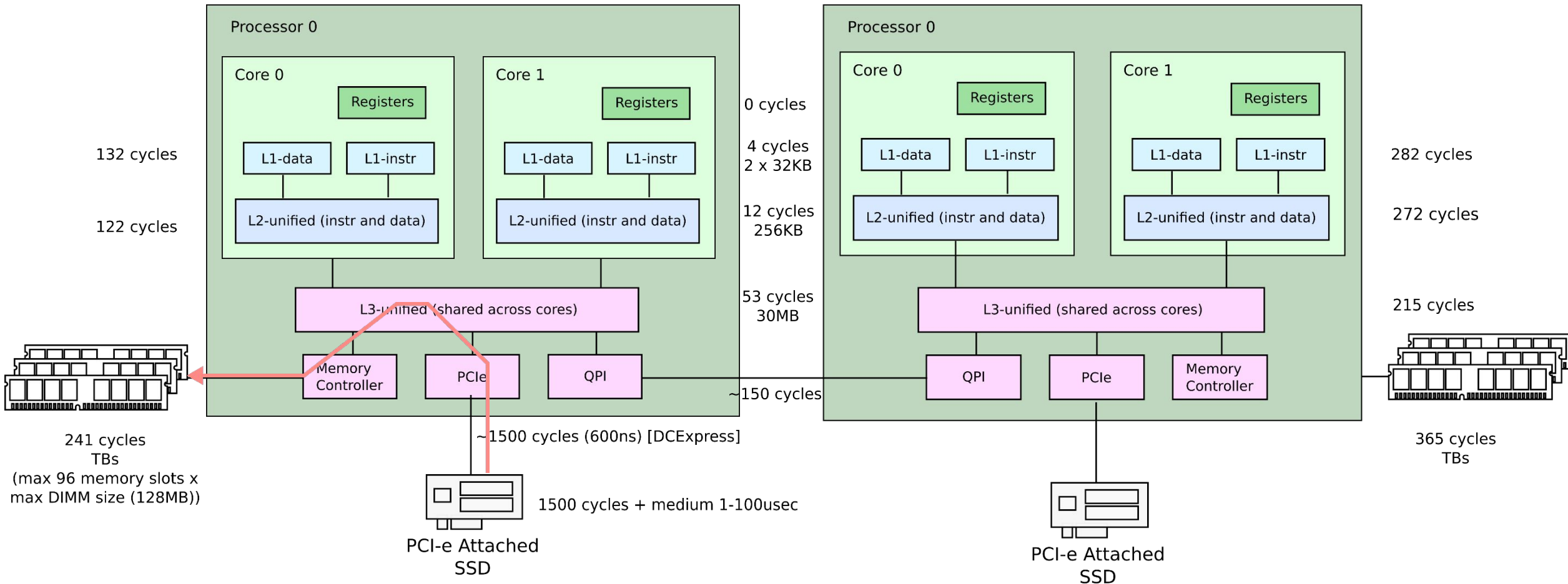
Latencies: PCIe round-trip



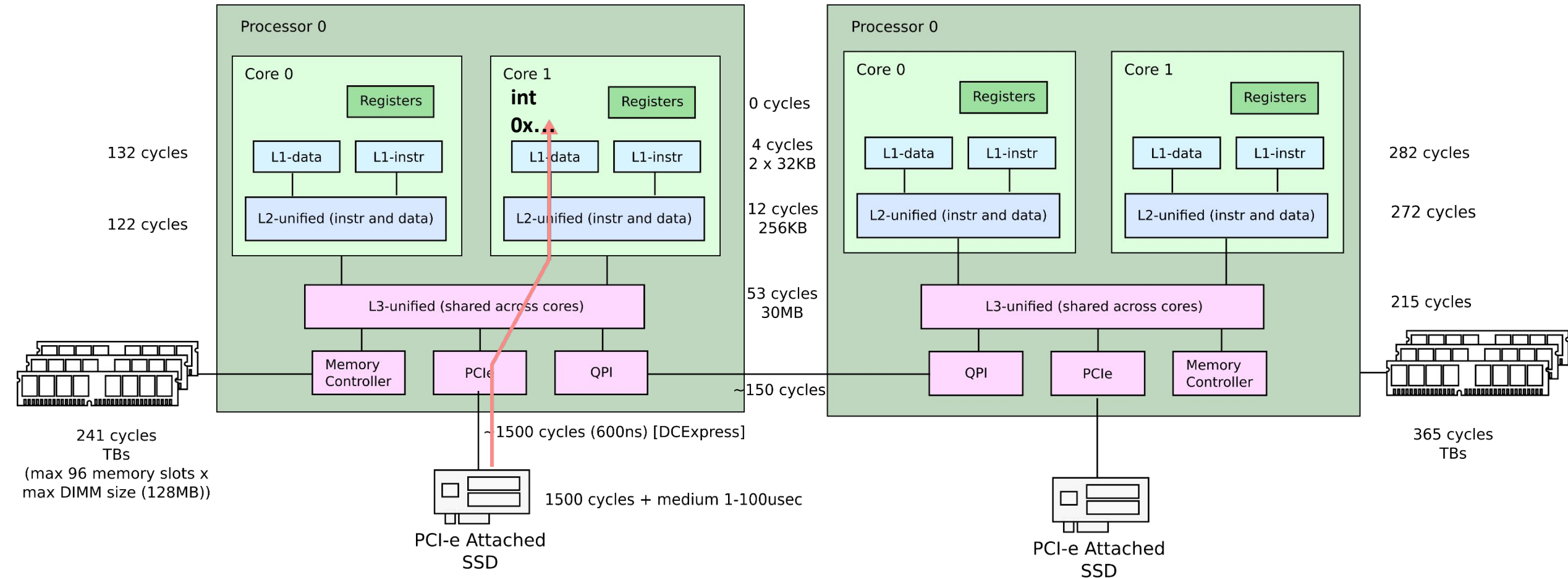
Device I/O

- Essentially just sending data to and from external devices
- Modern devices communicate over PCIe
 - Well there are other popular buses, e.g., USB, SATA (disks), etc.
 - Conceptually they are similar
- Devices can
 - Read memory
 - Send interrupts to the CPU

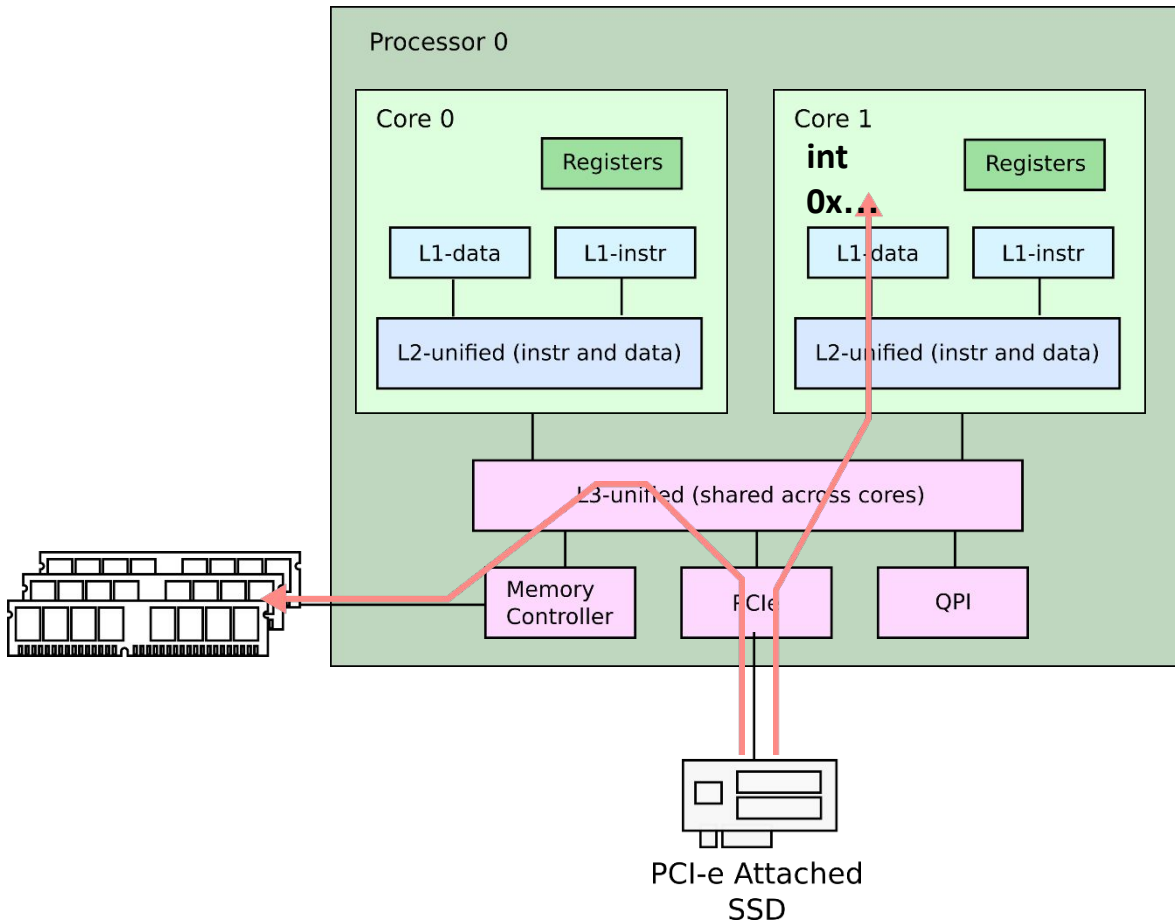
Direct memory access



Interrupts

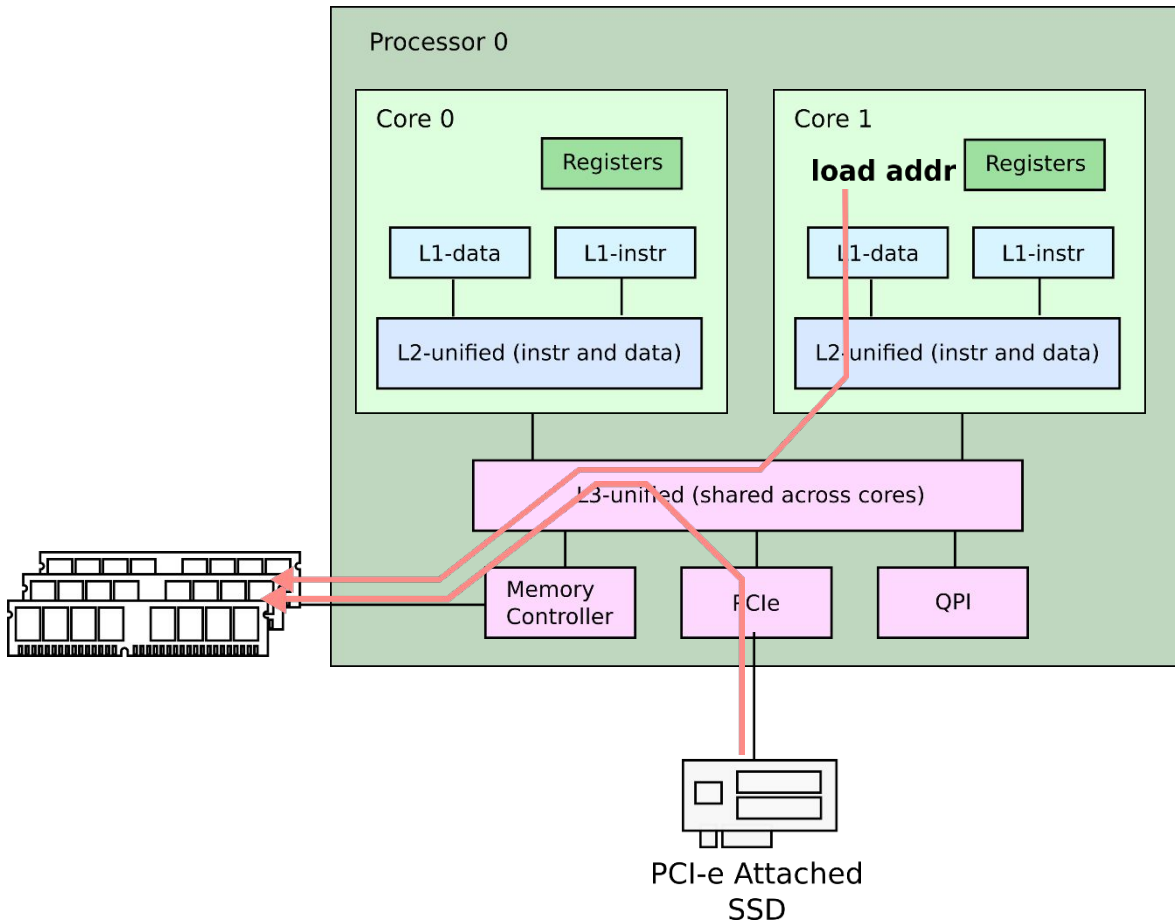


Device I/O



- Write incoming data in memory, e.g.,
 - Network packets
 - Disk requests, etc.
- Then raise an interrupt to notify the CPU
 - CPU starts executing interrupt handler
 - Then reads incoming packets from memory

Device I/O (polling mode)



- Alternatively the CPU has to check for incoming data in memory periodically
 - Or poll
- Rationale
 - Interrupts are expensive

References

- Cache Coherence Protocol and Memory Performance of the Intel Haswell-EP Architecture.
<http://ieeexplore.ieee.org/abstract/document/7349629>
- Intel SGX Explained <https://eprint.iacr.org/2016/086.pdf>
- DC Express: Shortest Latency Protocol for Reading Phase Change Memory over PCI Express
https://www.usenix.org/system/files/conference/fast14/fast14-paper_vucinic.pdf

Thank you!