SPARC : US4884198A US:-

Invention realized two important facts ,first to speedup the cache access and second to provision large cache capacity available to processor.It is well known fact that performance for majority of workloads including HPC, commercials like OLTP,DSS and web depends on the following three parameters.

a:-Reducing the miss rate.(means provisioning large cache is beneficial)
b:-Reducing the miss penalty.(Reducing the cache access time helps here)
c:-Reducing the time to hit in a cache(this contributes maximum since majority o accesses are hits, and it translates into performance)

Above optimizations are used even today to improve the performance for embedded caches like L3 over the generation of processors for example below T4 to T5 to M5/M6 the cache size and access optimizations :

L3 Cache Sizes(MB) :4MB(T4 Sparc),8MB(T5 sparc),48MB(M5,M6 sparc)
L3 cache latency (ns) 16.35ns(T4-Sparc) 14.44(T5 Sparc),14.47(M5,M6 sparc)

Above sizing has contributed to ~50% extra performance improvements on single thread workloads like spec2006 going from T4 to M5 considering frequency equivalence,higher cache sizes have also paid handsomely for JAVA and OLTP benchmarks to tune of 1,8x to 2.2x.

Motivation of this Patent was to optimize above three metrics to enhance IPC(instructions per second) using large external caches in Microprocessor based systems.This was particularly true when the large external cache presented a large capacitive load to the address bus of the processor. The driving of such large external, capacitive loads took cycles in the system and penalized the cache accesses time thus slowing down the program speed or instruction execution speed.

Invention realized the electrical physical limitations of technology(like capacitive loading of external cache) available at that time,delta in clock distribution delays to processor and to external component like a address register.It used these important facts to shave off the cache access time by starting cache accesses early(fraction of clock cycle) using a CAR(cache address register) than processor could have initiated with direct coupled address bus to cache.

The CAR also provided another benefit which was not feasible with a direct coupled processor address bus to cache,it was fabricated out of a technology that allowed it to drive the address to the large capacitive load of the cache memory in much less time than the processor itself could have driven such a load. Thus, due to this buffering capability of the CAR, the cache can be much larger compared to direct connect to processor itself.

Note that time expended sending the address from the processor to the CAR buffer, which would otherwise not be present if the processor addressed the cache directly from an internal register, does not subtract

from the processor cycle time since the processor can compute the cache address and send it to the CAR in less than the time required to access the cache.

In essence invention provided a method:
(l) utilizes inherent delays in the receipt of clock signals to provide additional time for cache access,thereby allowing for a shorter processor cycle time and a correspondent increase in the speed of instruction execution;

(2) allowed the cache to be larger than it otherwise could be for short cache access times even though the cache presented a large capacitive load to the memory address bus of the processor.