Abstract—Computer engineers are suffering in a hell partly of their own creation: while the time needed to access memory has slowly dropped, the number of instructions that can be executed during that time has increased exponentially. Since execution must sometimes wait for the outcome of a memory access, efforts at improving instruction execution speed are increasingly futile. This memory access latency can be partly avoided by caching data in a small, expensive, high-speed memory. Memory latency is still a problem because needed data is not always in the cache. An aggressive solution to the problem is prefetching: bringing data into the cache before it is needed using predictions of memory access addresses. Prefetch schemes reported in the literature predict consecutive or arithmetic address sequences, for example, 1000, 1025, 1050, 1075, etc. A more flexible scheme is described here: a neighborhood is found for memory access instructions based on their past behavior. Address predictions are made using a neighborhood, these can form sequential or stride sequences, but can also form irregular sequences, increasing the number of useful prefetches and reducing the number of useless ones. This neighborhood prefetching was evaluated by execution-driven simulation of shared-memory parallel computers (multiprocessors) running SPLASH 2 benchmark programs. Performance improvements of up to 25% were obtained on neighborhood prefetching schemes having storage requirements of only 7% that of the cache itself.