АПСТРАКТ

Од 1980 година развојот на процесорите и меморијата оди во две различни технолошки насоки што денес е причина за појава на јаз помеѓу нивните брзини. Таа разлика во брзините е причина за мемориската латентност кога процесорот пристапува до меморијата и тесно грло што ја блокира работата на процесорот[59]. Со развојот на процесорите и појавата на супер скаларни процесори се проширува нивниот инструкциски прозорец но тоа не е доволно за да ја прикрие мемориската латентност. Основната техника која се имплементира во процесорите за да се намали времето на пристап до меморијата е воведувањето на мемориската хиерархија[1],[2]. Меморискиот систем е поделен во неколку нива така што одејќи од процесорот кон основната меморија се зголемува капацитетот на кеш меморијата а се намалува брзината на пристап. Мемориската хиерархија ја намалува мемориската латентност но таа е причина за појава на кеш промашувања заради кои процесорот е потребно дополнително да губи време за нивно сервисирање. Истражувањата покажуваат дека кај современите процесори со повеќе јадра и нишки бројот на кеш промашувањата расте и ги турка перформансите на процесорот. Особено тоа е изразено при конкурентното пристапување кон споделената кеш меморија помеѓу нишките и јадрата. Тоа значи дека основен предизвик за истражувачите и дизајнерите во иднина е справувањето со кеш промашувањата.

Истражувањата од оваа област нудат повеќе техники за криење и намалување на мемориската латентност но ниедно од нив не е имплементирано од страна на дизајнерите во современите процесори. Основа во развојот на процесорите е односот цена-перформанси што значи дека е потребно да се дојде до техника која ќе ја намали мемориската латентност но е потребно да биде лесна за имплементација и нема дополнително да ја зголеми дисипацијата на енергија на процесорот.

Целта на ова истражување е да даде целосен преглед на досегашните истражувања во областа на паралелизам на мемориско ниво. Со симулатор на кеш мемориски систем се утврдува падот на перформансите на процесорот во зависност од кеш промашувањата кај современите процесори. Во понатамошното истражување е дизајниран симулатор на процесор користејќи

го пакетот *SimPy* и *Python* програмскиот јазик. Идејата е преку симулирањето на работата на различни типови на процесори да се дојде до мерливи резултати преку кои ќе може да се споредува ефикасноста на техниките кои се користат во намалување на мемориската латентност. Во докторската дисертација се моделирани симулатор на идеален процесор, процесор со предвремено преземање на мемориските податоци, процесор со имплементирана *runahead execution* техника, процесор со повеќе нишки. Анализата на добиените резултати и поведението на системот процесор-меморија дава идеја за воведување на техника кај современите процесори која може да ги намали конкурентните кеш промашувања и да ја намали латентноста на меморијата а со тоа да ги зголеми перформансите на процесорот за 15%.

Основниот придонес на ова научно истражувачко дело е:

1. Собирање на постојните истражувања од областа на паралелизмот на мемориско ниво и влијанието на кеш промашувањата врз перформансите на процесорот. Преглед на постојните техники теоретски и практични техники за намалување и криење на мемориската латентност.
2. Експериментално утврдување на влијанието на кеш промашувањата врз перформансите на современите процесори. Утврдување на параметрите кои и како влијаат врз бројот на кеш промашувањата.
3. Дизајнирање на симулатор на процесор кој овозможува да се следи поведението на системот процесор – меморија. Преку добиените резултати може да се врши анализирање на перформансите на различни типови на процесори кои во себе имаат имплементирано техники за намалување на латентноста. Во симулаторот лесно може да се имплементираат модули на нови техники при што преку резултатите да се споредува нивната ефикасност во однос на постојните типови на процесори. За прв пат во оваа докторска дисертација е дизајниран симулатор на процесор преку кого може да се тестира поведението на системот просецор-меморија и да се добијат мерливи споредбени резултати.

4. Како последен придонес е техниката која во докторската дисертација е предложена за зголемување на перформансите на процесорот со намалување на бројот на конкурентните кеш промашувања и намалување на латентноста на меморијата. Техниката ја отстранува конкурентноста во првото ниво на кеш меморијата од страна на нишките на јадрото на процесорот. Тоа би овозможило во иднина во процесорското јадро да се имплементираат повеќе нишки кои меѓусебно нема да ги истиснуваат мемориските податоци од кеш меморијата. Со тоа бројот на кеш промашувањата е намален за 15%. Од друга страна техниката врши предвремено преземање на мемориски податоци но во второто ниво на кеш меморијата што придонесува да се намали времето на пристап до меморијата и забрзување на извршувањето на програмите за 15%.

Клучни зборови: процесор, паралелизам на мемориско ниво, меморија, кеш меморија, мемориска хиерархија, кеш промашување, симулација на процесор, процесор со нишки

ABSTRACT

Starting from the 1980 the development of the processors and the memory splits into two different technological directions, which from today's perspective is the reason for the appearance of a gap between their speeds. This speed difference is the cause for memory latency when the processor accesses the memory and represents a bottleneck that blocks the work of the processor [59]. With the development of processors and the emergence of superscalar processors their instruction window is being expanded, but that isn't enough to conceal the latency of the memory. The basic technique implemented in the processors for lessening the memory access time is the introduction of memory hierarchy [1], [2]. The memory system is divided into several levels in such way that when going from the processor to the main memory the capacity of the cache memory is increasing and access speed is decreasing. Memory hierarchy reduces memory latency, but also is the reason for the appearance of cache misses for which the processor needs to waste extra time for their service. The research shows that in the modern processors with multiple cores and threads, the number of cache misses grows and decreases the processor performance. That is especially articulate at competitive access to the shared cache memory between the cores and the threads. That means that main challenge for scientists and designers in the future is tackling with cache misses.

The researches in this area offer more techniques for concealing and reducing memory latency, but designers in modern processors implement none of those. The basis in the development of processors is the price-performances ratio, which means that it is necessary to come to a technique that will reduce the memory latency, but at the same time to be easy to implement and not to increase the dissipation of processor power.

The aim of this research is to give a complete overview of current research in the area of parallelism at memory level. With a simulator of the cache memory system, the fall of the performances of a processor is determined depending on cache misses in modern processors. In the further research, a simulator of a processor is designed using the package SimPy and Python programming language. The idea is through simulation of different processor types to reach tangible results by which we can compare the effectiveness of the techniques used in reducing memory latency.

The dissertation modeled a simulator of an ideal processor, a processor with memory data prefetching, a processor with implemented runahead execution technique, and a multi-threaded processor. The analysis of performance and the behavior of the system processor-memory give an idea for the introduction of a technique in contemporary processors, which can lower cache misses and will reduce memory latency and increase processor performances for 15%.

The basic contribution of this scientific research work is:

1. Collection of existing research in the field of parallelism on memory level and the impact of cache misses on the processor performances. A review of existing techniques, both theoretical and practical for reducing and concealing memory latency.
2. Experimental determination of the impact of cache misses on modern processor performances. Determination of the parameters and how they influence the number of cache misses.
3. Designing a processor simulator, that allows monitoring of the behavior of system processor-memory. With the obtained results, the performance of different types of processors, which have implemented techniques for latency reducing, can be analyzed. In the simulator, we can easily implement modules of new techniques through which we can compare their efficiencies in comparison to existing types of processors. For the first time, in this dissertation a simulator of processor is designed, through which the behavior of the system, processor-memory can be tested and gain measurable comparable results.
4. As a last contribution in this dissertation is the technique, which is proposed to increase processor, performances with reducing the number of cache misses and memory latency. The technique removes the competitiveness in the first level of the cache memory by the threads in the core of the processor. This, in the future should enable more threads to be implemented in the processor core, which won't squeeze out each other's memory data from memory cache. With that, the number of cache misses will be reduced to 15%. In turn, the technique performs prefetching of data in the second

level of memory cache, which allows reducing of the memory access time and acceleration of the execution of programs for 15%.

Keywords: processor, memory level parallelism, memory, cache memory, memory hierarchy, cache misses