КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 3 С. 607–613

Ки&М

СЕКЦИОННЫЕ ДОКЛАДЫ

УДК: 004.27

Memory benchmarking characterisation of ARM-based SoCs

G. T. Wrigley[a], R. G. Reed, B. Mellado

School of Physics, University of the Witwatersrand. 1 Jan Smuts Avenue, Braamfontein, Johannesburg, South Africa, 2000

E-mail: [a] thomas.wrigley@cern.ch

Received September 30, 2014

Computational intensity is traditionally the focus of large-scale computing system designs, generally leaving such designs ill-equipped to efficiently handle throughput-oriented workloads. In addition, cost and energy consumption considerations for large-scale computing systems in general remain a source of concern. A potential solution involves using low-cost, low-power ARM processors in large arrays in a manner which provides massive parallelisation and high rates of data throughput (relative to existing large-scale computing designs). Giving greater priority to both throughput-rate and cost considerations increases the relevance of primary memory performance and design optimisations to overall system performance. Using several primary memory performance benchmarks to evaluate various aspects of RAM and cache performance, we provide characterisations of the performances of four different models of ARM-based system-on-chip, namely the Cortex-A9, Cortex-A7, Cortex-A15 r3p2 and Cortex-A15 r3p3. We then discuss the relevance of these results to high volume computing and the potential for ARM processors.

Keywords: ARM, memory, benchmarks, throughput-oriented computing, high-volume computing

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 3, pp. 607–617.

КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 3 С. 607–613

Ки&М

Описание тестирования памяти однокристальных систем на основе ARM

Г. Т. Ригли, Р. Г. Рид, Б. Мелладо

Отделение Физики, Университет Витватерсранда, Южная Африка, 2000, Йоханнесбург, 1 Ян Смут Авеню

Мощность вычислений традиционно находится в фокусе при разработке крупномасштабных вычислительных систем, в большинстве случаев такие проекты остаются плохо оборудованными и не могут эффективно справляться с ориентированными на высокую производительность рабочими нагрузками. Кроме того, стоимость и вопросы энергопотребления для крупномасштабных вычислительных систем всё ещё остаются источником беспокойства. Потенциальное решение включает в себя использование низко затратных процессоров ARM с маленькой мощностью в больших массивах в манере, которая обеспечивает массивное распараллеливание и высокую пропускную способность, производительность (относительно существующих крупномасштабных вычислительных проектов). Предоставление большего приоритета производительности и стоимости повышает значимость производительности оперативной памяти и оптимизации проекта до высокой производительности всей системы. Используя несколько эталонных тестов производительности оперативной памяти для оценки различных аспектов производительности RAM и кэш-памяти, мы даем описание производительности четырех различных моделей однокристальной системы на основе ARM, а именно Cortex-A9, Cortex-A7, Cortex-A15 r3p2 и Cortex-A15 r3p3. Затем мы обсуждаем значимость этих результатов для вычислений большого объема и потенциала для ARM- процессоров.

Ключевые слова: ARM-процессор, память, эталонные тесты, вычисления, ориентированные на высокую производительность, вычисления большого объема.

1. Introduction and Background

The volume of data generated by the wide array of available computing services in the consumer, industrial, academic and other spheres is vast and ever-increasing and the challenge posed by this is often called 'Big Data' — a term which is rapidly approaching ubiquity, with its wide array of potential applications generating a great deal of interest across many fields [Manyika et al., 2011]. Large-scale computing systems have traditionally been designed with computationally-intensive tasks as their primary focus. These systems are often highly inefficient for the purposes of throughput-oriented computing. A computing paradigm called High Volume Computing (HVC) has been proposed by Zhan et al [Zhan et al., 2012], which they define as a large number of loosely-coupled, throughput-oriented workloads, with increasing throughput volume being a principal goal of such system designs. A potential HVC solution involves the use of ARM processors, which are low-power, low-cost and low-energy consumption system-on-chips (SoCs), in large arrays which would provide very high levels of parallelisation. ARM-based SoCs, which are commonly used in mobile devices such as smartphones and tablets, are low-cost, mass-produced and potentially highly energy-efficient [Aroca, Gonçalves, 2012], all of which bodes well for both system affordability and energy efficiency. Although large scale computing has traditionally placed its primary focus on processor performance, there is an increasing shift towards including memory performance in this focus [Dongarra, Heroux, 2012; Ang et al., 2010]. Memory performance is a key component of overall system performance and is particularly important for throughput rates, memory bottlenecks could potentially affect energy-efficiency and cost through under-utilisation of existing system hardware. Using ARM-based SoCs in any proposed solution therefore requires that the performance of ARM-based SoCs be properly characterised and understood.

2. Experimental Configuration

The primary memory (i.e. RAM and cache) performance of four models of ARM SoC-based development boards were evaluated. Commercially available development boards were used for the purposes of benchmarking. The technical specifications of these boards are listed in Table 1 below.

Table 1. ARM development board hardware specifications

	Cortex-A7	**Cortex-A9**	**Cortex-A15 r2**	**Cortex-A15 r3**
Platform	Cubieboard2	Wandboard Quad	Odroid-XU+E	Jetson TK1
SoC	Allwinner A20	Freescale i.IMX6Q	Samsung Exynos 5410	NVIDIA Tegra K1
ARM Core Revision	r0p4	r2p2	r3p2	r3p3
Cores	2	4	4	4
Power-saver cores	0	0	4 Cortex-A7	1
Max. CPU Clock (MHz)	1008	996	1600	2300
L1 Cache (kB)	32	32	32	32
L2 Cache (kB)	256	1024	2048	2048
RAM Size (MB)	1024	2048	2048	2048
DDR3 RAM Type	432 MHz 32 bit	528 MHz 64 bit	800 MHz 64 bit	933 MHz 64bit DDR3L
Approx. 2014 Retail Price (USD)	65	129	169	192
Operating System	Ubuntu	Linaro	Ubuntu	Ubuntu

A Linux-based distribution was installed on all four board models. Three benchmarking software programmes were used to evaluate the memory performance of these four boards, namely the LMBench benchmark suite, the STREAM benchmark and the Parallel Memory Bandwidth Benchmark

(*pmbw*). The LMBench benchmarking suite analyses several aspects of memory performance — this study focuses on the measures of memory latency. The STREAM benchmark provides a measure of sustained memory bandwidth. STREAM works by generating an array of random numbers of a specified size (which is then stored in RAM) and performs four types of operations, namely copy, scale, add and triad. Measures of sustained bandwidth are then produced for each of these four tests. The *pmbw* benchmark is similar to STREAM in that it also provides a measure of memory bandwidth, but is also strongly influenced by memory latency. The *pmbw* benchmark consists of 14 separate subtests, each performing a slightly different operation. There are 5 variables which distinguish the 14 subtests, namely: (1) sequential scanning or a random access (permutation walking) test, (2) write or read operation, (3) bit size transferred in each operation, (4) pointer-based iterations vs index-based array access, and (5) number of operations per loop (1 — Simple vs 16 — Unroll) [Bingmann, 2013]. Two of the subtests involve Multiroll Loops and are not analysed here. The benchmark generates an array and runs one of the subtest routines. The allocated array size is then increased and the subtest routine is then repeated. This is repeated until the highest power of 2 able to fit onto the system's RAM is reached. These steps are repeated for each one of the subtest routines. *pmbw* is useful because it measures both bandwidth and latency and can potentially offer deeper insight into memory performance.

3. Results and Discussion

3.1. STREAM and LMBench

For the STREAM benchmark, which measures sustained memory bandwidth, the two Cortex-A15-based systems are clearly shown to be the best-performing of the four systems, with the r3p3 (Jetson TK1) obtaining the highest absolute bandwidth and the r3p2 (Odroid) obtaining the high bandwidth efficiency (i.e. percentage of theoretical maximum obtained). The Cortex-A7 displays reasonable bandwidth efficiency, while the Cortex-A9, which is the oldest of the four systems, achieves very low bandwidth efficiency, reaching only 16% of its theoretical maximum. In the case of RAM and cache latencies, the Cortex-A7, Cortex-A15p2 and Cortex-A15p3 all perform well, recording low latencies, with a clear correlation between CPU clock frequency and cache latency. The latency of the Cortex-A9 is also significantly higher the other three SoCs. For both of these benchmarks, a clear positive correlation can be seen between age of SoC design and performance. Table 2 below summarises the results obtained from both LMBench and STREAM for all four boards.

Table 2. LMBench and STREAM Benchmark Results

	Cortex-A7	**Cortex-A9**	**Cortex-A15 r2**	**Cortex-A15 r3**
Copy (MB/s)	1996	1329	6066	6430
Scale (MB/s)	1444	1110	6114	6403
Add (MB/s)	757	1448	5413	5358
Triad (MB/s)	702	1290	5275	5302
RAM (Theoretical MB/s)	3296	8054	12 207	14 236
Ave. RAM B/W Efficiency (%)	37	16	47	41
L1 Latency (ns)	3.02	4.02	2.51	1.73
L2 Latency (ns)	9.2	30.8	13.8	9.95
RAM Latency (ns)	58.5	119.8	104.8	115.6

3.2. The pmbw benchmark

The design of the *pmbw* benchmark means that each subtest routine generates several hundred sets of observations — between 200 and 300 observations in the case of the four systems tested here. Because there are several hundred observations per subtest and 12 subtests which are analysed here,

the volume of data produced by this benchmark for each system is very large — numbering around several thousand observations. For this reason, statistical tools are useful for extracting meaning from these data sets. A statistical test known as analysis of variance (ANOVA) was used to analyse the results of this benchmark. ANOVA is used to compare multiple datasets and determine whether the individual means of these datasets are equal to one another. More specifically, ANOVA compares the variance within each of these datasets to the variance which is present between these datasets and determines whether statistically significant differences exist between these datasets [Larson, 2008]. If statistically significant differences between these datasets do exist, various *post hoc* tests and analyses can then be used to gain greater insight into the distribution and nature of these differences.

In this case, each subtest (with its 200-300 observations per system) represents a dataset and ANOVA is used to determine whether these individual subtests are statistically similar to one another. A two-way analysis of variance showed that significant differences existed between the subtest groups for all four boards — i.e. at least one pair of means was different from one another. *Post hoc* analysis was then conducted to gain greater insight into the nature and distribution of these results. This analysis revealed the results generated by the 12 subtests appear to be distributed into five general groupings, with each grouping being made up of two, three or four subtests. As each subtest results from a combination of the benchmark's five function variables, the existence of these five groupings gives a greater level of insight into which of these characteristics appear to have the greatest impact on performance — insights which allow for memory performance to be better understood. The types of subtests which make up each grouping are briefly detailed in Table 3 below.

Table 3. Subtests contained *pmbw* in general result groupings

Group no.	Subtest types in group	Abbreviation
1	Random Pointer Permutations (Perm)	Random Pointer Permutation
2	Sequential Reading — 32 bit Simple Loop	SeqRead32Simple
3	Sequential Write — 32 bit Simple & Unroll Loop	SeqWrite32 Simple+Unroll
4	Sequential 32 bit Unroll & 64 bit Simple Loop	Seq32Unroll+64Simp
5	Sequential 64 bit Unroll Loop	Seq64Unroll

Based on the subtest result groupings determined above, the average of the two/three/four RAM bandwidth results for each of the five groupings was plotted. These bandwidth results are shown in Fig. 1 below. The first grouping (Random Pointer Permutation) is substantially lower than the other four groupings. This is, however, consistent with expectations, as this benchmark is based on a random pointer permutation and is essentially a measure of raw bandwidth and latency for one memory fetch cycle, while the other four are measures of sustained memory bandwidth for sequential scanning [Bingmann, 2013]. These results indicate that the Cortex-A7 (Cubieboard2) produces the lowest performance, while the Cortex-A9 (Wandboard) performing approximately 50% better. The Cortex-A15p2 (Odroid) performs approximately 50% better than the Wandboard, while the Cortex-A15p3 (Jetson TK1) in turn performs approximately 50% better than the Odroid. The Wandboard performing better than the Cubieboard2 appears to be inconsistent with the memory latency and sustained memory bandwidth results obtained by LMBench and STREAM, which showed the Cortex-A7 performing better than the Cortex-A9 in both cases. While these two random pointer permutation subtests are not solely dependent on memory latency, this would be expected to have some effect on random memory access performance. It is not immediately clear why the results produced by pmbw appear to conflict with the trends implied by the obtained LMBench results, although factors such as the Cortex-A9 SoC's 64 bit RAM bus width compared to the Cortex-A7 SoC's 32 bit RAM bus width may influence this result. This question must be further investigated in future work.

Groupings 2, 3, 4 and 5 are all based on sequential scanning rather than random memory access. This means that these four groupings offer some measure of sustained memory bandwidth. The very bandwidth measurements shown in Fig. 1 below are not directly comparable to STREAM as the measurements below represent average bandwidth, while the measurements in Tab. 2 for STREAM

are for sustained main memory (i.e. RAM) bandwidth. The general profile of the first three groups (i.e. Cortex-A7, Cortex-A9 & Cortex-A15p2) is consistent with the results obtained by the STREAM benchmark. The performance of the Cortex-A15p3 (NVIDIA Jetson TK1) is, however, more than twice that of the Cortex-A15p2 (Odroid-XU+E), while the Jetson TK1 only marginally outperforms the Odroid-XU+3 on the STREAM benchmark. This is most likely due to the influence higher clock frequency of the Jetson TK1 and its subsequently lower L1 and L2 cache latencies.

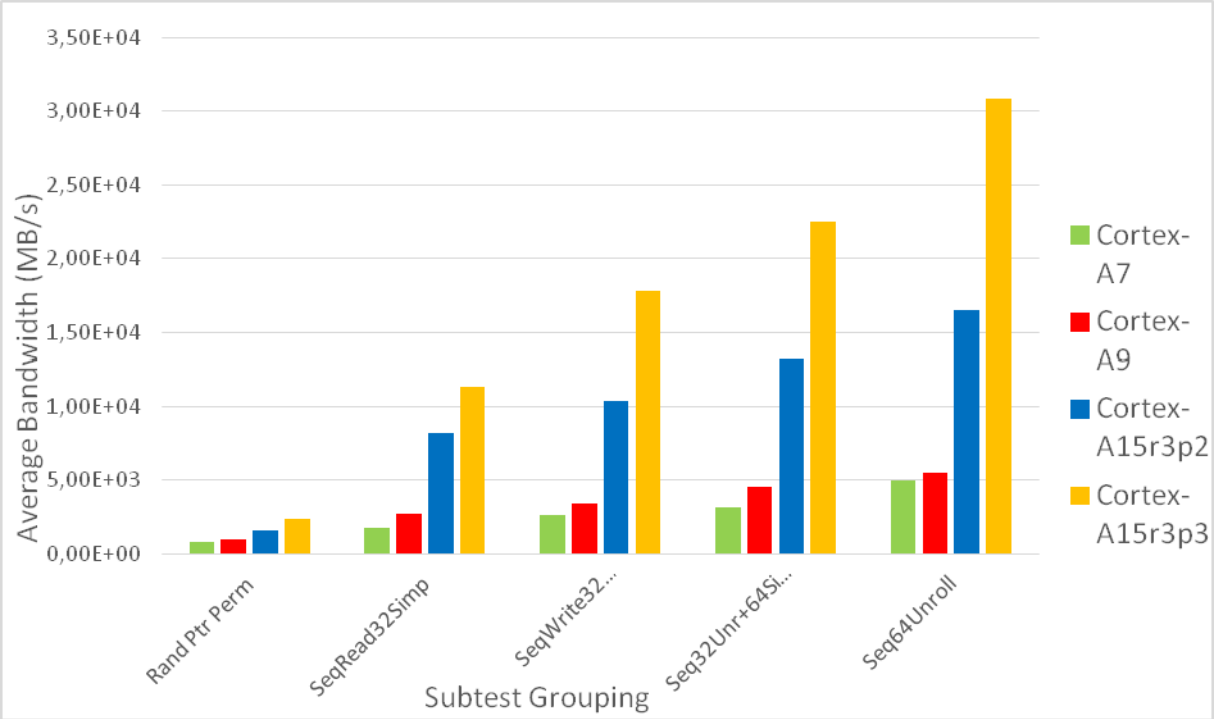

Fig. 1. *pmbw* Bandwidth Grouping Results

3.3. Discussion and Analysis

A clear correlation between age of SoC design and overall memory performance is clearly demonstrated, with the newest SoC, the Cortex-A15p3 performing the most effectively and the oldest SoC, the Cortex-A9 performing the least effectively. The bandwidth efficiency of the newest board (41%) is lower than the second-newest board (47%), which is something that can be improved upon in future board and SoC designs. Preliminary results presented at the 2014 South Africa Institute of Physics Conference by Mitchell Cox [Cox, 2014] show that it is possible to obtain I/O connection rates of approximately 300 MB/s between two Cortex-A9 SoCs. This suggests that memory performance is not the primary source of throughput rate bottlenecks for relatively simple algorithms (i.e. where CPU performance is not the bottleneck), as this figure is approximately 5 times lower than the sustained memory bandwidth measured for the Cortex-A9. As I/O connection rates continue to improve, this low sustained memory bandwidth may present an obstacle to overall throughput rates. The Cortex-A9 design tested here is, however, more than six years old. The performance improvements of the newer SoCs mean that I/O capacity is more likely to be the primary cause of throughput rate bottlenecks, particularly for algorithms which are not computationally intensive. These improvements are expected to continue as newer ARM-based SoCs are released, particularly with the soon-to-be released ARMv8 architecture 64 bit SoCs (such as NVIDIA's Project Denver). The potential of ARM-based SoCs for use in HVC systems therefore remains strong. Intel Atom-based SoCs hardware (i.e. development boards) will be procured in due course, in order to evaluate their potential for use in HVC.

Conclusion

In summary, the memory performance of four ARM SoC-based development boards was evaluated using three separate memory benchmarks. Of the four boards, the Cortex-A15r3p3 NVIDIA TK1 — the newest SoC design — was the best both in terms of sustained memory bandwidth and cache latency, reaching 6.4 GB/s for the former. Throughput-oriented workloads are thus unlikely to saturate memory, particularly for tasks which are more computationally complex than the STREAM benchmark. Although bandwidth efficiency for the newest board is lower than for the second-newest board, the general improvement in memory performance of the newer SoC designs displayed by these benchmarks suggest that memory performance will continue to improve in the near future. This suggests that ARM-based SoCs are viable candidates for use in HVC.

References

Ang, J. A., Barrett B. W., Wheeler K. B., Murphy R. C. 2010. Introducing the graph 500. No. SAND2010-3263C (Albuquerque, NM: Sandia National Laboratories).

Aroca R. V., Gonçalves L. M. G. Towards green data centers: A comparison of x86 and ARM architectures' power efficiency // Journal of Parallel and Distributed Computing. — 2012. — **72**. — P. 1770–1780.

Bingmann, T. pmbw — Parallel Memory Bandwidth Benchmark/Measurement. 2013. Retrieved from: http://panthema.net/2013/pmbw/

Cox, M. The development of a general purpose Processing Unit for the upgraded electronics of the ATLAS detector Tile Calorimeter, *SAIP 2014*. 2014 (submitted).

Dongarra, J., Heroux M. A. 2013. Toward a new metric for ranking high performance computing systems. No. SAND2013-4744 312 (Albuquerque, NM: Sandia National Laboratories).

Larson, M. G. Analysis of variance // Circulation. — 2008. — **117.1**. — P. 115–121.

Manyika, J., Chui M., Brown B., Bughin B., Dobbs R., Roxburgh C., Hung Byers A. 2011. Big data: The next frontier for innovation, competition and productivity. (New York City, NY: McKinsey Global Institute).

Zhan, J., Zhang L., Sun N., Wang L., Zhen J., Luo C. High volume throughput computing: Identifying and characterising throughput oriented workloads in data centers // Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012 IEEE 26th International. — 2012. — P. 1712–1721.